IMPERIAL

Imperial College London
Department of Mathematics

Change Point Detection for Extremal Graphical Models

James Cuin

CID: 02472833

Supervised by Yanbo Tang and Michaël Lalancette

30 August 2024

Submitted in partial fulfilment of the requirements for the MSc in Statistics at Imperial College London

The work contained in this thesis is my own work unless otherwise stated.

Signed: James Cuin Date: 30 August 2024

Acknowledgements

First and foremost, I would like express my immense gratitude to Yanbo Tang and Michaël Lalancette for their invaluable guidance and unwavering commitment throughout the development of this paper. I am truly grateful to them both for the exposition, and opportunity to contribute, in such a unique and dynamic subject area.

I am also incredibly grateful to my friends, and in particular to my housemates, for their patience and understanding during the long hours and late nights. I can only hope to repay their kindness.

Lastly, I would like to thank my family, particularly my parents, Mike and Cath, as well as my brother Adam, for their steadfast love and support, which ultimately made this paper possible.

Declarations

The data and code supporting the findings of this paper, as well as the explicit implementation of the proposed procedure, are openly accessible at https://github.com/jamescuin/egm-changepoint-detection. We acknowledge that simulation studies were conducted on the NextGen HPC cluster and MSc compute servers of Imperial College London, that the Information and Communication Technologies (ICT) department notably provided access, and to whom we are immensely thankful.

Abstract

The development of extremal graphical models, founded upon the notion of extremal conditional dependency, has facilitated sparse, interpretable modelling of such structural relationships. Under the popular modelling assumption of a Hüsler-Reiss distribution, significant progress pertaining to inference and consistent graph recovery has been achieved, yet the underlying model is notably assumed to be time-invariant. For scenarios in which this does not hold, specifically that in which piece-wise evolution of the underlying model is permitted, we leverage the connection between Hüsler-Reiss models and Gaussian graphical models by extending a change point detection methodology applicable to the latter. Focusing on the estimation of partition boundaries, we propose an extension of the regularised fused lasso procedure, whilst also outline the set of Karush-Kuhn-Tucker conditions necessarily satisfied by regression coefficients that recover these boundaries optimally. Through a suite of extensive simulation studies, the procedure is further shown to serve the dual purpose of screening and estimation, as is its inherent synergy with existing graph estimation methods through an application to currency exchange data.

Contents

1. Background & Introduction

Understanding extremal conditional dependence structures is paramount to effective quantification and mitigation of rare, yet catastrophic, events that are often predicated by simultaneous univariate rare events. On one hand, extreme value theory enables extrapolation to these univariate tails, whereas graphical models provide highly interpretable frameworks that facilitate sparse representations of the corresponding dependence structures, thereby enabling efficient utilisation of significantly smaller effective sample sizes. In the context of threshold exceedances, traditional conditional dependence is insufficient, as the support of multivariate Pareto distributions is contingent upon corresponding conditioning events, and thus a tailor-made notion of extremal conditional dependence is introduced (Engelke and Hitz, 2020), facilitating a natural marriage of the two fields. Under the Hüsler-Reiss model, the extremal analogue of the Gaussian graphical model, consistent recovery of the arbitrary underlying extremal graph is even possible (Engelke et al., 2024), although this notably assumes an absence of structural change points, which are particularly relevant in climate and financial systems.

In such cases, the assumption of identically distributed data is violated, necessitating the recovery of partition boundaries, between which relevant graph estimation methodologies should be applied, as otherwise recovered structures may represent a mixture of distributions, obscuring the true underlying dependencies at any given time. Although existing literature pertaining to the detection of structural change points in sequences of Gaussian graphical models is relatively sparse, the connection between Hüsler-Reiss and Gaussian distributions outlined in Section 3 renders this highly pertinent. Specifically, we focus on offline detection methods, and as we allow for multiple change points, the work of Kolar and Xing (2012) and Londschien et al. (2021) are of greatest relevance, although in high-dimensional settings Roy et al. (2017) offers an alternative approach, at the cost of estimating only a single change point. Indeed, both of the former papers focus on piece-wise evolution of the underlying structure, with the latter doing so in the presence of missing values, which while possible, are notably less prevalent in the extremal setting, due to the relative importance and infrequency of threshold exceedances.

Crucially, to the best of our knowledge, no formal literature related to that of structural change point detection in sequences of extremal graphical models exists, meaning that this work not only represents a truly novel methodology, of which is immediately applicable to any scenario that a Hüsler-Reiss graphical model is estimated, as in Manuel Hentschel and Segers (2024) and Engelke and Taeb (2024), but also highlights how, in general, one can extend Gaussian graphical model change point detection methodologies to the extremal setting. A comprehensive and compelling background to the sophisticated theory of relevant fields is provided in Sections 2 - 4, followed by the introduction of our proposed methodology in Section 5. A thorough suite of simulation studies, and motivating application, are then presented in Sections 6 and 7 respectively, whilst Section 8 provides a concluding discussion. Before proceeding, it is recommended that the notation schemes outlined in Appendix A are fully understood.

2. Graphical Models

Graphical models offer a compact and intuitive framework in which high-dimensional probability distributions can be understood from the viewpoint of dependency patterns, both causal and associative. Indeed, this framework is ubiquitous across not only statistical physics, in which Ising models, Kalman filters, and hidden Markov models are perhaps most relevant (Jordan, 2004), but also across other domains, such as bio-statistics and computer science (Lauritzen, 1996). Two distinct representations exist, namely that of undirected graphs, also known as Markov random fields, and directed (acyclic) graphs, also known as Bayesian networks, where we note, in the latter, the dependency structure is causal. Specifically, these dependency structures pertain to conditional independence, for which we refer to Maathuis et al. (2019) for a short, yet motivating, example.

2.1. Independence & Conditional Independence

We begin by revisiting the elementary, yet fundamental, concepts of independence and conditional independence for random variables, as avoiding conflation between the two will be of critical importance to our subsequent discussions.

Definition 2.1. Let d be a positive integer. A **product space** is the Cartesian product of individual sample spaces associated with each random variable. Formally, if $(\Omega_m, \mathcal{F}_m, \mathbb{P}_m)$ are probability spaces for $m \in \{1, \dots, d\} = [d]$, the product space is denoted by $(\Omega, \mathcal{F}, \mathbb{P})$, with

$$\begin{aligned} \Omega &= \Omega_1 \times \cdots \times \Omega_d, \\ \mathcal{F} &= \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_d, \\ \mathbb{P} &= \mathbb{P}_1 \times \cdots \times \mathbb{P}_d, \end{aligned}$$

where $\times$ denotes the Cartesian product and $\otimes$ denotes the product σ-algebra. In this space, each element is a tuple $(w_1, \dots, w_d)$, with $w_m \in \Omega_m$ for $m \in [d]$.

Indeed, this means that $\mathcal{F}$ is the smallest σ-algebra on Ω such that each coordinate projection, π_m, is measurable. Intuitively, this means that $\mathcal{F}$ contains all the sets required to ensure that π_m maps measurable sets in Ω_m to measurable sets in Ω. This concept, although commonly omitted, is crucial to our formulation of independence.

Definition 2.2. Given a finite collection of random variables, $X_1, \dots, X_d$, defined on a product space, $(\Omega, \mathcal{F}, \mathbb{P})$, we say that they are **independent** if, for all $x_1, \dots, x_d \in \mathbb{R}$,

$$F_{(X_1,\dots,X_d)}(x_1, \dots, x_d) = \prod_{m=1}^{d} F_{X_m}(x_m), \tag{1}$$

where $F_{(X_1,\dots,X_d)}$ is the $d.f.$ of $(X_1, \dots, X_d)$, and F_{X_m} is the marginal $d.f.$ of X_m.

Equivalently, for all Borel sets $B_1, \ldots, B_d \subseteq \mathbb{R}$,

$$\mathbb{P}\left(X_1 \in B_1, \ldots, X_d \in B_d\right) = \prod_{m=1}^{d} \mathbb{P}(X_m \in B_m). \tag{2}$$

The independence between two random variables, say X_m and $X_{m'}$, with $m \neq m'$, is denoted by $X_m \perp\!\!\!\perp X_{m'}$.

Importantly, this implies that the joint probability distribution can be factored into the product of marginal distributions over the entire product space. Indeed, in either case of this collection of random variables being discrete or continuous, we have that Equation 1 is equivalent to

$$f_{(X_1,\ldots,X_d)}\left(x_1, \ldots, x_d\right) = \prod_{m=1}^{d} f_{X_m}(x_m), \quad \forall x_1, \ldots, x_d \in \mathbb{R}, \tag{3}$$

where f_{X_m} denotes the probability mass function (for discrete random variables) or probability density function (for continuous random variables) of X_m.

In fact, the notion of independence relies on the assumption that the joint distribution is defined over a product space where each random variable can vary independently. When the support of this distribution is restricted, such as $\mathcal{L} = \left\{x \in \mathbb{R}^d : \|x\|_\infty > t,\ t > 0\right\}$ for example, this factorisation may no longer hold.

Now, conditional independence, which we define below, is naturally related to independence, and for a thorough treatment and outline of associated concepts, such as conditional expectation, we refer to Proschan and Shaw (2016).

Definition 2.3. Given a finite collection of random variables, $X_1, \ldots, X_d$, and another random variable Z, both defined on a product space, $(\Omega, \mathcal{F}, \mathbb{P})$, we say that $X_1, \ldots, X_d$ are **conditionally independent** given Z if, $\forall x_1, \ldots, x_d \in \mathbb{R}$ and $\forall z \in \mathbb{R}$,

$$\mathbb{P}\left(X_1 \leq x_1, \ldots, X_d \leq x_d \mid Z\right) = F_{(X_1,\ldots,X_d|Z)}\left(x_1, \ldots, x_d \mid z\right) = \prod_{m=1}^{d} F_{X_m|Z}(x_m \mid z), \tag{4}$$

where $F_{X_m|Z}$ is the conditional $d.f.$ of X_m given Z. The conditional independence between two random variables, say X_m and $X_{m'}$, with $m \neq m'$, given Z, is denoted by $X_m \perp\!\!\!\perp X_{m'} \mid Z$.

It is important to note that the definition of the conditional $d.f.$ for a discrete random variable differs to that for a continuous random variable, as in the latter case we have $\mathbb{P}(X_m = x_m) = 0$ for every $x_m \in \mathbb{R}$.

Similarly to the case of independence, for discrete or continuous random variables, we have that Equation 4 is equivalent to the following

$$f_{(X_1,\ldots,X_d|Z)}\left(x_1, \ldots, x_d \mid z\right) = \prod_{m=1}^{d} f_{X_m|Z}(x_m \mid z), \quad \forall x_1, \ldots, x_d \in \mathbb{R}, \forall z \in \mathbb{R}, \tag{5}$$

where $f_{X_m|Z}$ denotes the conditional probability mass or density function of X_m given Z.

Critically, conditional independence is similarly reliant on the product space structure, as the factorisation into the product of conditional distributions assumes that the joint distribution given Z is defined over the entire product space.

Example 2.1. Suppose X and Y are random variables with support $\mathcal{L}$, and we condition on the random variable Z. Well, if $\mathcal{L} = \{(x, y) \in \mathbb{R}^2 : \max(x, y) > t,\ t > 0\}$, then for any fixed $Z = z$, the support constraint still holds, and so X and Y cannot both be less than $t > 0$ simultaneously, meaning that the conditional joint density is zero in the region $[0, t]^2$. However, the conditional marginal densities are not necessarily zero in the regions $x \leq t$ and $y \leq t$ respectively, and so the factorisation, $f_{(X,Y|Z)}(x, y|z) = f_{X|Z}(x|z)f_{Y|Z}(y|z)$ does not hold over the entirety of $\mathbb{R}^2$.

2.2. Probabilistic Graphical Models

Now, we introduce the notion of a graph, as well as associated terminology, and subsequently outline how these can be utilised to determine conditional dependency structures.

Definition 2.4. A **graph** is characterised through the pair $\mathcal{G} = (V, E)$, where $V = [d]$ is a finite set of *nodes*, commonly referred to as vertices, and $E \subset V \times V$ is a set of *edges*, that consists of ordered pairs of distinct nodes, denoted by $(\alpha, \beta) \in E$, where $\alpha, \beta \in V$.
(Lauritzen, 1996)

We remark that the graphs considered in the following are *simple*, in that no loops from a node to itself, for example $(\alpha, \alpha) \in E$, exist, nor do multiple edges between two nodes. In fact, we will differentiate between two common graphical representations of distributions, that of Bayesian networks and Markov random fields, where we note the former is based upon directed (acyclic) graphs, and the latter on undirected graphs.

Definition 2.5. The pair of nodes, $(\alpha, \beta) \in E$, are said to be connected by a **directed edge** if $(\beta, \alpha) \notin E$, denoted by $\alpha \to \beta$. If, however, we have $(\alpha, \beta) \in E$ and $(\beta, \alpha) \in E$, then the nodes are said to be connected by an **undirected edge**, denoted by $\alpha - \beta$. If a graph consists of all directed edges, or of all undirected edges, then the graph is said to be **directed** and **undirected** respectively.

Definition 2.6. The nodes, α to β, form a **path** of length k, in $\mathcal{G}$, if $\alpha_0 = \alpha, \ldots, \alpha_k = \beta$ construct a sequence of distinct nodes such that $\alpha_{m-1} \to \alpha_m$ or $\alpha_{m-1} - \alpha_m$, for all $m \in [k]$.
(Lauritzen, 1996)

Definition 2.7. A **k-cycle**, in $\mathcal{G}$, is a path of length k, $\alpha_0, \ldots, \alpha_k$, where $\alpha_0 = \alpha_k$, that is the path begins and ends at the same node. The cycle is said to be directed if it contains a directed edge. A graph is **acyclic** if it contains no cycles. (Lauritzen, 1996)

Definition 2.8. A graph is known as a **directed acyclic graph** (DAG) if it is both directed and acyclic.

Indeed, probabilistic graphical models (PGMs) leverage the aforementioned graph based representation to encode either independencies, or factorisation, in a highly compact manner, which, as seen in Section 2.1, were shown to be equivalent. Both Bayesian networks and Markov random fields are examples of PGMs, each of which can be utilised to encode a differing set of independencies (Koller and Friedman, 2009). Bayesian networks use DAGs to represent asymmetric factorisations of joint probability distributions, while Markov random fields utilise undirected graphs (UGs) to capture symmetric relationships. Given the inherent symmetry of conditional independence, we focus on the latter, particularly on UGs, for which the structure is fully defined through unordered pairs of nodes, $\{\alpha, \beta\}$, reflecting this symmetry.

Before proceeding, we first introduce a variety of Markov properties for UGs, which we refer to as graphs from now on, as these provide a natural way to interpret graphical models. Indeed, we now explicitly consider a collection of random variables, $(X_m)_{m \in V}$, which take values in the probability spaces $(\mathcal{X}_m)_{m \in V}$, as is outlined in Lauritzen (1996), where recall $\mathcal{X}_m = (\Omega_m, \mathcal{F}_m, \mathbb{P}_m)$. To be clear, we denote the product space by $\mathcal{X} = \mathcal{X}_V = \times_{m \in V} \mathcal{X}_m$.

Definition 2.9. The probability distribution on $\mathcal{X}$ is said to satisfy the **pairwise Markov property**, with respect to the graph, $\mathcal{G} = (V, E)$, if, for any pair (α, β) of non-adjacent nodes,

$$(\alpha, \beta) \notin E \;\Rightarrow\; X_\alpha \perp\!\!\!\perp X_\beta | \mathbf{X}_{\setminus \{\alpha, \beta\}}.$$

Before outlining the local and global Markov properties, however, we require notions of *neighbourhoods* and *separation* within a graph, $\mathcal{G}$.

Definition 2.10. The **neighbourhood**, of node α, is given by $N(\alpha) = \{\beta \in V_{\setminus \alpha} : (\alpha, \beta) \in E\}$.

Definition 2.11. If $A, B, C \subset V$ are disjoint subsets, then we say that C **separates** A from B in $\mathcal{G}$ if every path from a node in A to a node in B contains a node in C.

Definition 2.12. The probability distribution on $\mathcal{X}$ is said to satisfy the **local Markov property**, with respect to the graph, $\mathcal{G} = (V, E)$, if, for any node $\alpha \in V$,

$$(\alpha, \beta) \notin E \;\Rightarrow\; X_\alpha \perp\!\!\!\perp X_\beta | \mathbf{X}_{N(\alpha)},$$

where $N(\alpha)$ is the neighbourhood of node α.

Definition 2.13. The probability distribution on $\mathcal{X}$ is said to satisfy the **global Markov property**, with respect to the graph, $\mathcal{G} = (V, E)$, if, for any three disjoint subsets, $A, B, C \subset V$, we have

$$C \text{ separates } A \text{ and } B \;\Rightarrow\; \mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B | \mathbf{X}_C.$$

It is important to recognise that certain Markov properties are stronger than others. In fact, we have that the global Markov property ⇒ local Markov property ⇒ pairwise Markov property, which we refer to Lauritzen (1996) for the proof. Critical to our discussion, however, is the fact that if the joint density, $f_{\mathbf{X}}(\mathbf{x})$, of the random vector, $\mathbf{X} = (X_1, \ldots, X_d)$, is strictly larger than zero, then in fact we have the pairwise Markov property ⇒ global Markov property (Ripley, 1996), and so equivalency between all properties holds. Indeed, this fact is utilised as part of the Hammersley-Clifford theorem, which we introduce in our subsequent discussion.

First, we outline the concept of *factorisation* in graphs, and highlight its connection to the Markov properties, and note that we saw a similar relationship for conditional independence.

Definition 2.14. A graph, $\mathcal{G} = (V, E)$, is said to be **complete**, if $\forall \alpha, \beta \in V$, with $\alpha \neq \beta$, we have $(\alpha, \beta) \in E$, that is each distinct node pair is connected by an edge. If this holds for a sub-graph, $\mathcal{G}_S \subset \mathcal{G}$, then we refer to $\mathcal{G}_S$ as a **complete sub-graph**.

Definition 2.15. A complete sub-graph of $\mathcal{G}$ that is not contained within any other complete sub-graph is known as a **maximal clique**, which we refer to as a **clique** from now on. By "contained", we mean that the clique cannot be extended by an additional adjacent node.

Consequently, a clique of size k, which we refer to as a k-clique, has ${}^kC_2$ undirected edges, where C is the binomial coefficient.

Definition 2.16. For a graph, $\mathcal{G} = (V, E)$, a probability measure, $\mathbb{P}$ on $\mathcal{X}$, is said to **factorise** on $\mathcal{G}$, if for all complete sub-graphs, A, there exist non-negative functions, ϕ_A, such that the joint density is characterised by

$$f_{\mathbf{X}}(\boldsymbol{x}) = \prod_A \phi_A(\boldsymbol{x}_A). \tag{6}$$

Note that the ϕ_A are not uniquely determined, and can be split-up or multiplied together, but must be suitable functions on $\times_{a \in A} \mathcal{X}_a$.

Indeed, as a clique is a maximal complete sub-graph, we could instead consider only these cliques for the factorisation, giving the following formulation,

$$f_{\mathbf{X}}(\boldsymbol{x}) = \prod_{C \in \mathcal{C}} \phi_C(\boldsymbol{x}_C), \tag{7}$$

where $\mathcal{C}$ denotes the set of cliques in $\mathcal{G}$. (Lauritzen, 1996)

Theorem 2.1. (Hammersely-Clifford) If the joint density is continuous, such that $f_{\mathbf{X}}(\boldsymbol{x}) > 0$, then the product measure, μ on $\mathcal{X}$, factorises on $\mathcal{G}$ if and only if the product measure satisfies the pairwise Markov property. Furthermore, as $f_{\mathbf{X}}(\boldsymbol{x}) > 0$, the pairwise, local, and global Markov properties are all equivalent.

Indeed, the Hammersely-Clifford Theorem establishes a crucial equivalence between the pairwise Markov property and the factorisation of the joint density function. Specifically, if the pairwise Markov property holds, then we can factorise the joint density into a product of functions defined on cliques, which significantly reduces the complexity of dealing with high-dimensional distributions, to (multiple) lower-dimensional ones. Although we do not currently have a natural way to specify the functions ϕ_C, we remark that if the graph, $\mathcal{G}$, is additionally decomposable, we are able to reformulate the factorisation in terms of the marginal distributions over the cliques.

Definition 2.17. A graph, $\mathcal{G} = (V, E)$, is said to be **decomposable** if either:

- $\mathcal{G}$ is complete, or
- $V = A \cup B \cup C$, with A and C non-empty, and B separates A from C, where B is complete, and both the resulting sub-graphs are themselves decomposable.

In fact, $\mathcal{G}$ is decomposable if and only if the set of cliques, $\mathcal{C} = \{C_1, \ldots, C_l\}$, are such that, for all $i = 2, \ldots, l$,

$$S_i = C_i \cap \bigcup_{j=1}^{i-1} C_j \subset C_k, \quad \text{for some } k < i, \tag{8}$$

which is known as the running intersection property (Green and Thomas, 2013). The S_i are known as **separators** of $\mathcal{G}$, and are uniquely determined for a given ordering of cliques.

Proposition 2.1. If the graph, $\mathcal{G} = (V, E)$, is additionally decomposable, then the factorisation condition outlined in Equation 7 can be re-written in terms of the product of marginal distributions over the cliques, divided by the product of the distributions on their intersections,

$$f_{\mathbf{X}}(\boldsymbol{x}) = \frac{\prod_{C \in \mathcal{C}} f_C(\boldsymbol{x}_C)}{\prod_{S \in \mathcal{S}} f_S(\boldsymbol{x}_S)}, \tag{9}$$

where the $S \in \mathcal{S}$ are separators of $\mathcal{G}$.

2.3. Gaussian Graphical Models

Markov random fields (MRFs) take on either discrete or continuous forms, where a notable example for the former is the Ising model which is famously used to model magnetic field interactions. In the continuous case, the random vector, $\mathbf{X} \in \mathbb{R}^d$, associated with the network

is assumed to follow a continuous distribution, such as the multivariate Gaussian distribution, in the case of a Gaussian graphical model (GGM).

To be clear, for a GGM, we have that $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the subscript d highlights the dimension of the distribution, and where we assume $\boldsymbol{\Sigma}$ is positive definite. The inverse of the covariance matrix, $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, is known as the *precision matrix*, and so we can write the joint density as follows:

$$\begin{aligned} f_{\mathbf{X}}(\boldsymbol{x}) &= \frac{(\det\{\boldsymbol{\Theta}\})^{\frac{1}{2}}}{(2\pi)^{\frac{d}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Theta}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \\ &= \frac{(\det\{\boldsymbol{\Theta}\})^{\frac{1}{2}}}{(2\pi)^{\frac{d}{2}}} \prod_{m,l} \exp\left\{-\frac{1}{2}(x_m-\mu_m)^{\top}\Theta_{ml}(x_l-\mu_l)\right\}. \end{aligned}$$

We remark that the above is a factorisation of $f_{\mathbf{X}}(\boldsymbol{x})$, which note is a strictly positive and continuous density. Thus, by Theorem 2.1, we can conclude that the pairwise Markov property is satisfied, and so the absence of edges in the associated graph indicates this form of conditional independence. Furthermore, we note this factorisation is specifically for $\Theta_{ml} \neq 0$, as when $\Theta_{ml} = 0$ we have that the exponent term is given by $\exp\{0\} = 1$, and so becomes a constant term that does not represent any meaningful interaction between the variables. In this sense, we have that $\boldsymbol{\Theta}$ encodes the conditional dependency structure of $\mathbf{X}$, and specifically for the graph, $\mathcal{G} = (V, E)$, we have that there exists an edge, $(\alpha, \beta) \in E$, if $\Theta_{\alpha\beta} \neq 0$. To be clear, if $\Theta_{\alpha\beta} = 0$, then $X_\alpha \perp\!\!\!\perp X_\beta | X_{\setminus\{\alpha,\beta\}}$.

A notable property of GGMs is stability under marginalisation, in that when we marginalise over a subset of the nodes, say A, the resulting distribution associated with this subset is still a GGM. Indeed, this follows from the fact that the marginal distribution of any subset, of a multivariate Gaussian distribution, is also a multivariate Gaussian distribution, and consequently the relevant submatrix of $\boldsymbol{\Theta}$ still encodes the relevant conditional dependency structure of $\mathbf{X}_A$. Importantly, if $\mathcal{G}$ is decomposable, then $f_{\mathbf{X}}(\boldsymbol{x})$ can be re-written as outlined in Proposition 2.1, where $f_C(\boldsymbol{x}_C)$ and $f_S(\boldsymbol{x}_S)$ are lower-dimensional Gaussian distributions.

Initially, it may be tempting to estimate $\boldsymbol{\Sigma}$ by the sample covariance, $\hat{\boldsymbol{\Sigma}}$, however, to recover an estimation of $\mathcal{G}$, we require an estimation of $\boldsymbol{\Theta}$, which is ideally sparse. Indeed, this would require the invertibility of $\boldsymbol{\Sigma}$ to hold, which we remark is not the case when $n < d$, that is when the sample covariance matrix is singular. Consequently, various regularisation algorithms have been proposed to transform $\hat{\boldsymbol{\Sigma}}$ into an estimate of the zero pattern of $\boldsymbol{\Theta}$, without this requirement of invertibility. These algorithms, denoted by $\mathcal{B}$, are commonly referred to as *base learners*, of which the most commonly utilised are that of neighborhood selection (Meinshausen and Bühlmann, 2006) and graphical lasso (Friedman et al., 2007). The former method can intuitively be thought of as node-wise l_1-regularised regression, and has $\mathcal{O}(nd^2 \min(n, d))$ time complexity for n realisations of the random vector. On the other hand, the latter method, which is effectively l_1-regularised maximum likelihood estimation, has a time complexity of $\mathcal{O}(nd^3)$ and is thus potentially far slower than that of neighborhood selection.

Example 2.2. Consider the GGMs outlined in Figure 1, where we denote the graphs before and after the change point by $\mathcal{G}_b$ and $\mathcal{G}_a$ respectively. In the case of $\mathcal{G}_b$, the cliques and separators, $\mathcal{C}_b$ and S_b, are given by:

$$\mathcal{C}_b = (\{1,2\},\{2,3\},\{2,5\},\{3,4\},\{4,5\},\{4,6\}), \quad S_b = (\{2\},\{4\})$$

whereas for $\mathcal{G}_a$, the cliques and separators, $\mathcal{C}_a$ and S_a, are given by:

$$\mathcal{C}_a = (\{1,2,3\},\{2,3,4,5\},\{4,5,6\}), \quad S_a = (\{2,3\},\{4,5\})$$

Due to the presence of the 4-cycle without a chord, that is a connection between non-adjacent nodes in the cycle, $\mathcal{G}_a$ is non-decomposable. Indeed, $\mathcal{G}_a$ is neither complete, nor can it be decomposed such that a complete sub-graph separates two non-empty sub-graphs. On the other hand, $\mathcal{G}_b$ is decomposable, as the clique $\{2,3,4,5\}$ is a complete sub-graph that separates $\{1\}$ and $\{6\}$. Note, $\mathbf{\Theta}_b$ and $\mathbf{\Theta}_a$ necessarily have the following zero patterns:

$$\mathbf{\Theta}_b = \begin{pmatrix} * & * & 0 & 0 & 0 & 0 \\ * & * & * & 0 & * & 0 \\ 0 & * & * & * & 0 & 0 \\ 0 & 0 & * & * & * & * \\ 0 & * & 0 & * & * & 0 \\ 0 & 0 & 0 & * & 0 & * \end{pmatrix} \qquad \mathbf{\Theta}_a = \begin{pmatrix} * & * & * & 0 & 0 & 0 \\ * & * & * & * & * & 0 \\ * & * & * & * & * & 0 \\ 0 & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & 0 & 0 & * & * & * \end{pmatrix}$$

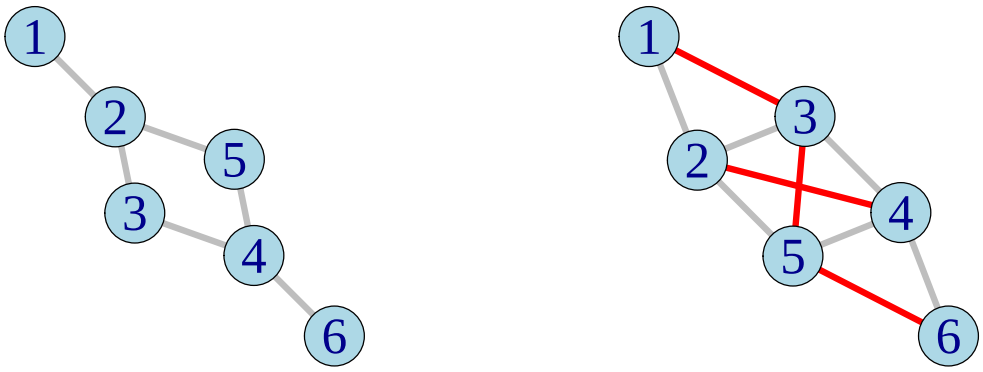

Figure 1.: A $d = 6$ GGM, before (left) and after (right) a change point. New conditional dependence relationships are indicated by edges highlighted in red.

3. Extreme Value Theory

Two distinct approaches exist for modelling tail behaviour of random variables and random vectors, one based upon componentwise maxima, and the other upon threshold exceedance. Although the following presentation pertains to maxima, note that since $\min(x_i) = -\max(-x_i)$, one can straightforwardly reformulate these approaches for minima also.

3.1. Block Maxima

For clarity, we begin by considering the univariate case, that is independent and identically distributed (i.i.d) random variables, $X_1, \ldots, X_n$, with underlying $d.f.$ denoted by F. In contrast to the central limit theorem (CLT), which is concerned with the limiting distribution of the average, we instead aim to understand the limiting distribution of the maxima of these random variables.

Indeed, if we let $M_n = \max\{X_1, \ldots, X_n\} = \max_{i=1}^{n} X_i$, we have:

$$\mathbb{P}(M_n \leq x) = \mathbb{P}(X_1 \leq x, \ldots, X_n \leq x) = F^n(x)$$

which note converges, as $n \to \infty$, to a degenerate limit distribution, specifically to 0 for $x \leq x_F$, and to 1 for $x \geq x_F$, where $x_F = \sup\{x \in \mathbb{R} : F(x) \leq 1\} \in (0, \infty]$ denotes the right endpoint of F. Thus, to find a non-degenerate limit distribution, the concept of the maximum domain of attraction is required, where similarly to the CLT, we utilise a location-scale normalisation.

Definition 3.1. The distribution function, F, is in the **maximum domain of attraction** (MDA) of distribution function, H, if $\exists\,(c_n)_{n\in\mathbb{N}} \in (0,\infty), (d_n)_{n\in\mathbb{N}} \in \mathbb{R}$, such that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{M_n - d_n}{c_n} \leq x\right) = \lim_{n\to\infty} F^n(c_n x + d_n) = H(x), \tag{10}$$

and H is non-degenerate. This is often written as $F \in \mathrm{MDA}(H)$.

In fact, as outlined in Theorem 3.1 (Fisher and Tippett, 1928), this non-degenerate limit must be a generalised extreme value distribution.

Definition 3.2. The **generalised extreme value** (GEV) distribution, with $\xi \in \mathbb{R}, \mu \in \mathbb{R}$, and $\sigma > 0$, satisfies

$$H_{\xi,\mu,\sigma}(x) = \begin{cases} \exp\left(-\left(1 + \xi\frac{x-\mu}{\sigma}\right)_+^{-\frac{1}{\xi}}\right), & \text{if } \xi \neq 0, \\ \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right), & \text{if } \xi = 0, \end{cases} \tag{11}$$

where $z_+ = \max(0, z)$.

Theorem 3.1. (Fisher-Tippett, Gnedenko). If $F \in \mathrm{MDA}(H)$, then H is a member of the GEV family, denoted by $H_{\xi,\mu,\sigma}(x)$, with $\xi \in \mathbb{R}, \mu \in \mathbb{R}$, and $\sigma > 0$.

In fact, GEV distributions are equivalent to the so-called max-stable distributions.

Definition 3.3. The distribution function, F, is called **max-stable** if $\exists\ c_n \in (0,\infty), d_n \in \mathbb{R}$, for every $n > 1$, such that

$$F^n(c_n x + d_n) = F(x), \tag{12}$$

where recall that F^n is the $d.f.$ of M_n.

Theorem 3.2. A $d.f.$ is max-stable if and only if it is a GEV. (Coles, 2001)

In the multivariate setting, we instead consider i.i.d random vectors, $\mathbf{X}_i = (X_{i,1}, \cdots, X_{i,d})$, where $i \in [n]$, that are d-dimensional, with $d.f.$, F. Furthermore, we denote the componentwise maxima by $M_{n,m} = \max_{i=1}^{n} X_{i,m}$, where $m \in [d]$, and the collection by $\mathbf{M}_n = (M_{n,1}, \ldots, M_{n,d})$.

Then, if $\exists\ (c_{n,m})_{n\in\mathbb{N}} \in (0,\infty), (d_{n,m})_{n\in\mathbb{N}} \in \mathbb{R}\ \forall\ m \in [d]$, such that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\mathbf{M}_n - \mathbf{d}_n}{\mathbf{c}_n} \leq \boldsymbol{x}\right) = \lim_{n\to\infty} F^n(\mathbf{c}_n \boldsymbol{x} + \mathbf{d}_n) = G(\boldsymbol{x}), \tag{13}$$

we say that F is in the multivariate MDA of G, where the latter $d.f.$ is referred to as the multivariate extreme value distribution. To be clear, note that $F(\boldsymbol{x}) = (F_1(x_1), \ldots, F_d(x_d))$. Indeed, the univariate margins of G are GEV distributions (Fougères, 2003), which are thus also max-stable by Theorem 3.2, and can be transformed, under certain location and scale corrections, into one of the three classical extreme-value distributions (Kotz and Nadarajah, 2000).

As outlined in Bücher and Zhou (2021), the block maxima method (Gumbel, 1958) involves taking only the maxima of each block, thus posing an issue relating to data efficiency. Indeed, not only do we expect other extreme observations to be informative of marginal and joint tail behaviour in general, but also that the clustering of these extremes to be of great importance, particularly in the context of change point detection. It is, however, the fact that the aforementioned max-stable distributions do not lead to meaningful notions of (extremal) conditional independence (Engelke and Hitz, 2020), that motivates the use of multivariate threshold exceedances. Specifically, if a max-stable admits a positive continuous density, then the pairwise Markov property implies joint independence, resulting in uninteresting Markov structures (Papastathopoulos and Strokorb, 2016).

3.2. Threshold Exceedances

There are, in fact, two approaches related to the threshold exceedances method, often referred to as the peaks-over-threshold (POT) method, namely the fully parametric models derived from the Generalised Pareto distribution, and the semi-parametric models that rely on the Hill Estimator (Abad et al., 2014). We focus on the former, and begin by considering the univariate case.

As the method name suggests, it is natural to consider observations that exceed a specified threshold, u, as extreme. Indeed, it is the distribution of these exceedances, rather than the maxima, that is now of interest.

Definition 3.4. The **distribution of exceedances** of X, over the threshold, $u < x_F$, is defined as

$$F_u(x) = \mathbb{P}\left(X - u \leq x | X > u\right) = \frac{F(x+u) - F(u)}{1 - F(u)}, \quad x \in [0, x_F - u]. \tag{14}$$

Importantly, note that if F were known, then by Definition 3.4, $F_u(x)$ is necessarily known. This of course is rarely the case, which recall motivated the characterisation of the limiting distribution for F, and thus we now seek a limiting distribution for $F_u(x)$, however note that now we consider the limit of $u \to x_F$, rather than $n \to \infty$. Indeed, Theorem 3.3 can be seen as the counterpart of Theorem 3.1 for threshold exceedances.

Definition 3.5. The **generalised Pareto distribution** (GPD) has the d.f.,

$$G_{\xi,a(u)}(x) = \begin{cases} 1 - \left(1 + \frac{\xi x}{a(u)}\right)^{-\frac{1}{\xi}}, & \text{if } \xi \neq 0, \\ 1 - \exp\left(-\frac{x}{a(u)}\right), & \text{if } \xi = 0, \end{cases} \tag{15}$$

where $a(u) = \sigma + \xi(u - \mu)$ and we require that $x \geq 0$ if $\xi \geq 0$, and $0 \leq x \leq -a(u)/\xi$ if $\xi < 0$.

Theorem 3.3. (Pickands-Balkema-de Haan). Let $G_{\xi,a(u)}(x)$ be a generalised Pareto distribution with $\xi \in \mathbb{R}$ and $a > 0$. Then, there is a positive, measurable function, $u \to a(u)$, such that

$$\lim_{u \to x_F} \sup_{x \in [0, x_F - u]} |F_u(x) - G_{\xi,a(u)}(x)| = 0, \tag{16}$$

if and only if F satisfies Theorem 3.1.

As for the GEV distributions, special cases of the GPD result in perhaps more familiar distributions, where clearly the $\xi = 0$ case is equivalent to an exponential distribution with rate $\lambda = 1/a(u)$, whereas the $\xi = -1$ case results in a continuous uniform distribution, $U\left(0, a(u)\right)$. Indeed, asymptotic equivalence exists between the block maxima and POT approaches (Vignotto and Engelke, 2020), however we now turn to the multivariate case in the interest of brevity.

In the multivariate setting, we effectively utilise the definition of a multivariate Pareto (MP) outlined in Rootzén and Tajvidi (2006), although we present this in a similar manner to Engelke and Volgushev (2022) below, as it is a far more intuitive account. To this end, we begin by considering observations, $\mathbf{X}_i$, such that $F(\mathbf{X}_i) \not\leq 1 - q$, as extreme, that is the observations such that one or more component of $\mathbf{X}_i$ exceeds some margin quantile, $1 - q$. Importantly, to

ensure that the limiting distribution of these extremes is guaranteed to exist, we require that a random vector, $\mathbf{Y}_i$, with support given by

$$\mathcal{L} = [0, \infty)^d \setminus [0, 1]^d, \tag{17}$$

exists, and that for all continuity points, $\mathbf{x} \in \mathcal{L}$, of the $d.f.$ of $\mathbf{Y}_i$, we have

$$\mathbb{P}\left(\mathbf{Y}_i \leq \mathbf{x}\right) = \lim_{q \to 0} \mathbb{P}\left(F(\mathbf{X}_i) \leq 1 - \frac{q}{\mathbf{x}} \mid F(\mathbf{X}_i) \not\leq 1 - q\right). \tag{18}$$

Intuitively, this means that if at least one component of $\mathbf{X}_i$ exceeds the margin quantile, $1 - q$, then we have that $q/(1 - F(\mathbf{X}_i)) \xrightarrow{\mathrm{D}} \mathbf{Y}_i$. Indeed, as outlined in Engelke and Volgushev (2022), this requirement arises from a regularity condition known as multivariate regular variation, of which a detailed treatment is provided in de Fondeville and Davison (2018). We note that practical details, pertaining to the selection of margin quantile, are discussed in Appendix D.

Definition 3.6. Let the limit in Equation 18 hold. Then, the random vector, $\mathbf{Y}_i \in \mathbb{R}^d$, follows a **multivariate Pareto distribution** (MPD), and the following properties hold,

1. $\mathbf{Y}_i \in \mathcal{L} = [0, \infty)^d \setminus [0, 1]^d$,
2. $\mathbb{P}\left(Y_{i,1} > 1\right) = \cdots = \mathbb{P}\left(Y_{i,d} > 1\right)$,
3. $\mathbb{P}\left(\mathbf{Y}_i \in tA\right) = t^{-1}\mathbb{P}\left(\mathbf{Y}_i \in A\right)$, where $A \subset \mathcal{L}$ and $t \geq 1$.

Furthermore, the random vector, $\mathbf{X}_i$ is said to be in the **Pareto max-domain of attraction** of the MPD, $\mathbf{Y}_i$.

Inspecting the homogeneity property illuminates how the MPD is related to the Pareto distribution. Indeed, we have that

$$\mathbb{P}\left(Y_{i,m} > x\right) = \mathbb{P}\left(Y_{i,m} \in x[1, \infty)\right) = x^{-1}\mathbb{P}\left(Y_{i,m} \in [1, \infty)\right) = x^{-1}\mathbb{P}\left(Y_{i,m} > 1\right),$$

and so we have

$$\begin{aligned}\mathbb{P}\left(Y_{i,m} \leq x \mid Y_{i,m} > 1\right) = 1 - \mathbb{P}\left(Y_{i,m} > x \mid Y_{i,m} > 1\right) &= 1 - x^{-1}\mathbb{P}\left(Y_{i,m} > 1 \mid Y_{i,m} > 1\right) \\ &= 1 - x^{-1},\end{aligned}$$

which is the standard Pareto distribution, with shape parameter $\alpha = 1$.

The class of MPDs is rich, including parametric families such as the extremal Dirichlet and logistic distributions, as well as others, including those considered in Cooley et al. (2010) and de Carvalho and Davison (2014). Furthermore, we note that an admittedly more technical presentation of certain aspects of the two approaches for modelling tail behaviour seen above, in terms of the exponent measure, a type of radon measure, can be found in Engelke and Hitz (2020), whilst key details pertaining to this measure are outlined in Appendix C.

3.3. Hüsler-Reiss Distributions

The d-dimensional Hüsler-Reiss distribution (Hüsler and Reiss, 1989) is a MPD, and as outlined shortly, can be considered as the counterpart to the Gaussian distribution in the extremal setting. In light of the absence of similar transformations to that presented in Lemma 3.1, for other MPDs, this distribution is of utmost importance to our discussion. Various presentations of this distribution exist, see Gudendorf and Segers (2010) and Lalancette (2023), however we opt for the following.

Definition 3.7. The d-dimensional **Hüsler-Reiss** (HR) distribution is a MPD parameterised by a variogram matrix, $\Gamma \in \mathcal{D}_d$, where

$$\mathcal{D}_d = \left\{ \Gamma \in \mathbb{R}^{d \times d} : \Gamma = \Gamma^\top,\ \operatorname{diag}(\Gamma) = \mathbf{0},\ v^\top \Gamma v < 0\ \forall\ \mathbf{0} \neq v \perp \mathbf{1} \right\}. \tag{19}$$

Indeed, if $\mathbf{Y} \sim \mathrm{HR}(\Gamma)$, where $\mathbf{Y} \in \mathbb{R}^d$, then for any root node, $m \in [d]$, we have

$$f_{\mathbf{Y}}(\boldsymbol{y}) \propto y_m^{-2} \left(\prod_{i \neq m} y_i^{-1} \right) \varphi_{d-1} \left\{ \log\left(\frac{\boldsymbol{y}_{\setminus m}}{y_m} \right) + \frac{\Gamma_{\setminus m,m}}{2},\ \Sigma^{(m)} \right\}, \tag{20}$$

where $\varphi_{d-1}\{\cdot,\ \Sigma^{(m)}\}$ is the $(d-1)$ normal density, with covariance matrix,

$$\Sigma^{(m)} = \frac{1}{2}(\Gamma_{im} + \Gamma_{jm} - \Gamma_{ij})_{i,j \neq m} \in \mathbb{R}^{d-1 \times d-1}. \tag{21}$$

Although perhaps not immediately apparent, we note that the above characterisation does not, in fact, depend on the choice of m.

An important quantity for HR distributions, and in fact for MPDs in general, are the auxiliary random vectors, $\mathbf{Y}^{(m)}$, of $\mathbf{Y} \in \mathbb{R}^d$, defined as $\mathbf{Y}^{(m)} = (\mathbf{Y} \mid Y_m > 1)$, where $m \in [d]$. In fact, these allow us to introduce the extremal variogram, a common extremal summary statistic that diverges to ∞ whenever Y_i and Y_j become asymptotically independent, although we note that the reverse direction does not necessarily hold.

Definition 3.8. Let $\mathbf{Y} \in \mathbb{R}^d$ be a MPD. Then, we define the **extremal variogram** rooted at node, $m \in [d]$, as the matrix $\Gamma^{(m)}$, such that

$$\Gamma_{ij}^{(m)} = \operatorname{Var}\left\{ \log\left(\mathbf{Y}_i^{(m)} \right) - \log\left(\mathbf{Y}_j^{(m)} \right) \right\}, \quad i, j \in [d], \tag{22}$$

provided the variance exists and is finite. (Engelke and Volgushev, 2022)

In the case of the HR distribution, the extremal variogram, rooted at any node $m \in [d]$, is equivalent to the variogram matrix, Γ (Engelke et al., 2024), and so we have that

$$\Gamma = \Gamma^{(1)} = \cdots = \Gamma^{(d)},$$

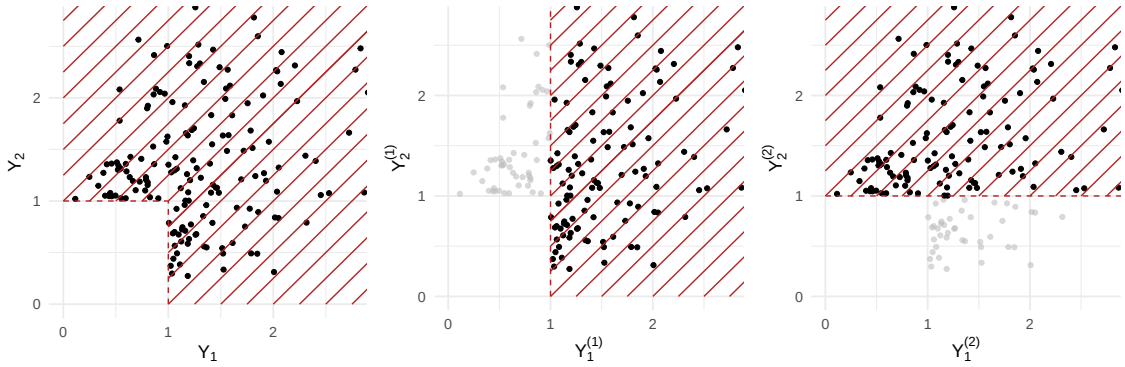

Figure 2.: Exact samples for $\mathbf{Y} \sim \mathrm{HR}_2(\Gamma)$, with the support for $\mathbf{Y}$, $\mathbf{Y}^{(1)}$, and $\mathbf{Y}^{(2)}$, from left to right, indicated by the red cross hatchings. Samples lost through the conditioning of the respective auxiliary random vector are faded.

meaning the extremal variograms will be of particular relevance to subsequent discussions, and in the case that Γ is unknown, natural estimates, such as the average of the empirical extremal variograms, which under mild assumptions satisfies concentration bounds outlined in Engelke et al. (2024), can be leveraged.

Another quantity of particular importance, which note is defined through the extremal variogram, is the precision matrix, $\Theta^{(m)} = \left(\Sigma^{(m)}\right)^{-1}$, that not only allows us to re-write Equation 20 as follows,

$$f_{\mathbf{Y}}(\boldsymbol{y}) \propto y_m^{-2} \left(\prod_{i \neq m} y_i^{-1} \right) \exp\left\{ -\frac{1}{2} \mathbf{z}^\top \Theta^{(m)} \mathbf{z} \right\}, \tag{23}$$

where $\mathbf{z} = \log(\boldsymbol{y}) - \log(y_m)\,\mathbf{1} + \Gamma_{\setminus m,m}/2$, but also encodes the extremal dependence structure, which we introduce in Section 4, through its sparsity pattern. Indeed, it is exactly this precision matrix that is estimated in the `EGlearn` majority voting algorithm (Engelke et al., 2024), which notably utilises one of the aforementioned base learners.

Lemma 3.1. (Gaussian Transformation) Let $\mathbf{Y} \sim \mathrm{HR}(\Gamma)$, where $\mathbf{Y} \in \mathbb{R}^d$, then

$$\tilde{\mathbf{Y}}_{(m)} = \log\left(\mathbf{Y}_{\setminus m}^{(m)}\right) - \log\left(\mathbf{Y}_{m}^{(m)}\right) \sim \mathcal{N}_{d-1}\left(-\frac{\mathrm{diag}\left(\Sigma^{(m)}\right)}{2},\ \Sigma^{(m)} \right). \tag{24}$$

Proof. See Appendix B.1 □

Critically, given a HR distributed random vector, $\mathbf{Y} \in \mathbb{R}^d$, we can leverage Lemma 3.1 to obtain d, one for each node, instances of a $(d-1)$-dimensional multivariate Gaussian distributed random vector, with mean vector and covariance matrix fully determined by the extremal variogram, Γ. Indeed, note how in Figure 5, $\mathbf{Y} \in \mathbb{R}^2$ results in two sets of 1-dimensional random vectors that exhibit a Gaussian distribution.

4. Extremal Graphical Models

As previously outlined, of critical importance to the notion of conditional independence is that of the product space upon which the random vector, $\mathbf{Y} \in \mathbb{R}^d$, is defined over. In particular, note that the support of MPDs, $\mathcal{L}$, as given by Property 1 in Definition 3.6, is a special case of the support outlined in Example 2.1, with $t = 1$, and is thus not a product space. Consequently, a new notion of conditional independence for MPDs is required, namely that of extremal conditional independence, which, as outlined in the groundbreaking work of Engelke and Hitz (2020), utilises auxiliary random vectors to restrict the random vector within a product space, specifically within $\mathcal{L}$.

Definition 4.1. The m-th **auxiliary random vector** of the random vector $\mathbf{Y} \in \mathbb{R}^d$, which we denote $\mathbf{Y}^{(m)}$, is given by $\mathbf{Y}^{(m)} = (\mathbf{Y} \mid Y_m > 1)$, where $m \in [d]$.

Definition 4.2. Let $\mathbf{Y} \in \mathbb{R}^d$ be a MPD that admits a positive and continuous joint density, $f_{\mathbf{Y}}$, on its support, $\mathcal{L}$. Furthermore, let $V = A \cup B \cup C$, where A, B, C are non-empty disjoint subsets, and $V = [d]$. If we have that

$$\mathbf{Y}_A^{(m)} \perp\!\!\!\perp \mathbf{Y}_B^{(m)} \mid \mathbf{Y}_C^{(m)}, \quad \forall m \in V, \tag{25}$$

where $\mathbf{Y}_S^{(m)}$ is the auxiliary random vector associated with set S, then we say that $\mathbf{Y}_A$ has **extremal conditional independence** of $\mathbf{Y}_B$ given $\mathbf{Y}_C$. To differentiate between extremal conditional independence and regular conditional independence, we denote the former by $\mathbf{Y}_A^{(m)} \perp_e \mathbf{Y}_B^{(m)} \mid \mathbf{Y}_C^{(m)}$. (Engelke and Hitz, 2020)

Intuitively, therefore, extremal conditional independence can be understood as regular conditional independence holding in the m different restrictions of $\mathcal{L}$, such that each restriction guarantees a product space within $\mathcal{L}$. A more subtle point is that the requirement of a positive and continuous joint density is not strictly necessary, and is required for the factorisation of densities as seen for the exponent measure in Engelke and Hitz (2020), whilst also preventing a natural extension for extremal independence. Indeed, this leads to the conclusion that if graphical models are formulated in regards to Definition 4.2 of extremal conditional independence, then they must be necessarily connected (Engelke and Hitz, 2020). If, however, as suggested and proved in Strokorb (2020), we remove this requirement, then a natural extension for extremal independence exists, where $B = \emptyset$, and is equivalent to the regular asymptotic independence. In this case, the associated graph may be disconnected. Regarding the associated exponent measure, Λ, extremal independence holds for $\mathbf{Y}_A$ and $\mathbf{Y}_B$ if mass is placed only on the faces $\mathcal{E}^S = \left\{\mathbf{Y} \in \mathcal{E} : \boldsymbol{y}_S > \mathbf{0}, \boldsymbol{y}_{\setminus S} = \mathbf{0}\right\}$, where S is a subset of either A or B (Strokorb, 2020).

Now, we have all the components to formally introduce the notion of an extremal graphical model, where the phrase 'extremal' specifically refers to threshold exceedances.

Definition 4.3. Let $\mathcal{G} = (V, E)$ be a graph with a finite set of nodes, $V = [d]$, and let $\mathbf{Y} \in \mathbb{R}^d$ be a MPD. If $\mathbf{Y}$ satisfies the (extremal) pairwise Markov property on $\mathcal{L}$ with respect to $\mathcal{G}$,

$$(\alpha, \beta) \notin E \;\Rightarrow\; Y_\alpha \perp_e Y_\beta \mid \mathbf{Y}_{\setminus\{\alpha,\beta\}}, \tag{26}$$

then we say that the distribution of $\mathbf{Y}$ is an **extremal graphical model** with respect to $\mathcal{G}$. (Engelke and Hitz, 2020)

4.1. Hüsler-Reiss Graphical Models

Similar to GGMs in the non-extremal setting, Hüsler-Reiss Graphical Models (HRGMs) exhibit desirable properties pertaining to stability and compact conditional dependency representation that make them natural models to consider in the extremal setting. To be clear, for a HRGM, we have that $\mathbf{Y} \sim \mathrm{HR}_d(\Gamma)$, where the subscript d again highlights the dimension of the distribution, and $\Gamma \in \mathcal{D}_d$.

Indeed, the extremal conditional dependency structure, and thus the extremal graph structure, in light of Definition 4.3, is encoded through the precision matrix, $\Theta^{(m)}$, specifically through the absence of zeros. This is formalised in Proposition 4.1.

Proposition 4.1. Let $\mathbf{Y} \sim \mathrm{HR}_d(\Gamma)$, and $\alpha, \beta \in [d]$, such that $\alpha \neq \beta$. Then, for any $m \in [d]$, we have that

$$Y_\alpha \perp_e Y_\beta \mid \mathbf{Y}_{\setminus\{\alpha,\beta\}} \;\Leftrightarrow\; \begin{cases} \Theta^{(m)}_{\alpha\beta} = 0, & \text{if } \alpha, \beta \neq m, \\ \sum_{l \neq m} \theta^{(m)}_{l\beta} = 0, & \text{if } \alpha = m, \beta \neq m, \\ \sum_{l \neq m} \theta^{(m)}_{\alpha l} = 0, & \text{if } \alpha \neq m, \beta = m. \end{cases} \tag{27}$$

(Engelke and Hitz, 2020)

Proof. See Appendix F.6 of Engelke and Hitz (2020). □

Furthermore, we remark that, for MPDs with positive and continuous densities in general, given a decomposable graph, $\mathcal{G} = (V, E)$, which we recall must be connected, a comparable theorem to that of Theorem 2.1, and factorisation to that outlined in Proposition 2.1, exists (Engelke and Hitz, 2020). Since a d-dimensional HR distribution is also stable under marginalisation, where specifically for a subset $A \subset V$, and any $m \in A$, we have that

$$f_{\mathbf{Y}_A}(\boldsymbol{y}_A) \propto y_m^{-2} \left(\prod_{i \in A \setminus \{m\}} y_i^{-1} \right) \varphi_{|A|-1} \left\{ -\frac{1}{2} \mathbf{z}^\top_{A\setminus\{m\}} \Theta^{(m)}_A \mathbf{z}_{A\setminus\{m\}},\ \Sigma^{(m)}_A \right\}, \tag{28}$$

we also have that $f_{\mathbf{Y}}(\mathbf{y})$ can be re-written in terms of lower-dimensional HR distributions.

5. Model Development

As previously outlined, no formal literature pertaining to that of structural change point detection for sequences of extremal graphical models exists, posing a significant challenge, yet opportunity, to fill this gap. Importantly, if only HRGMs are considered, then we can, in light of Lemma 3.1, transform the d-dimensional HRGM into d instances of $(d-1)$-dimensional GGMs, and subsequently leverage relevant change point detection methodology for the latter type of graphical model. Furthermore, this Lemma can be viewed as a property of the HR distribution, as acknowledged in Engelke and Hitz (2020) and Röttger et al. (2023), however we are unaware of any methodology explicitly leveraging this fact, and thus the subsequent methodology further represents a novel technical application of said property. Although existing work for change point detection in GGMs is admittedly sparse, as outlined in Section 1, they do importantly provide general frameworks and essential intuition that we adapt and extend for the extremal setting.

Specifically, as this is the first formal attempt in detecting structural change points for extremal graphical models, we build upon the work of Kolar and Xing (2012), which addresses the simpler case of piece-wise evolution in the underlying precision matrix, with notably no missing data. Indeed, in the case of GGMs, the covariance matrix, $\boldsymbol{\Sigma}$, fully determines the precision matrix, $\boldsymbol{\Theta}$, whilst the covariance matrices of the resulting GGMs obtained through Lemma 3.1, $\Sigma^{(m)}$, are determined solely by the extremal variogram matrix, Γ, of the respective HR distribution. Thus, the task of detecting piece-wise changes in the underlying extremal variogram matrix, which recall characterises the HRGM, can be reformulated to that of detecting piece-wise changes in d precision matrices. In fact, this is the foundation for our proposed methodology, that we now formally outline below, along with the problem setup.

5.1. Problem Setup

Consider a sequence of independent HR random vectors, $(\boldsymbol{y}_i)_{i\in[n]} \in \mathbb{R}^d$, with extremal variograms, Γ_i, such that the map $i \mapsto \Gamma_i$ is piece-wise constant. We aim to estimate, for an unknown number of change points, N_C, the set of change points, $\mathcal{T} = \{t_1, \ldots, t_{N_C}\}$, that form a disjoint partitioning of $[n]$, such that

$$\mathcal{T} = \{i : \Gamma_i \neq \Gamma_{i+1}\}, \tag{29}$$

where these change points are naturally ordered, in that $t_1 < \cdots < t_{N_C}$. To be clear, the absence of change points, that is $N_C = 0$, is certainly valid, and in this case $\mathcal{T} = \emptyset$. Furthermore, note that a change point, therefore, can correspond to either an explicit change in the structure of the HRGM, that is a change in the extremal conditional dependency structure, or simply a change in the strength of the extremal conditional dependencies.

5.2. Proposed Procedure

As alluded to, the first, and perhaps most crucial, step of the proposed procedure consists of leveraging Lemma 3.1 to transform the $(\boldsymbol{y}_i)_{i\in[n]} \in \mathbb{R}^d$ into d sets of multivariate Gaussian distributed random vectors, where we denote the m-th set by $\left(\tilde{\boldsymbol{y}}_{i'}^{(m)}\right)_{i'\in[n^{(m)}]} \in \mathbb{R}^{d-1}$, and we have $n^{(m)} \leq n$. Indeed, the aforementioned transformation, which we subsequently denote by φ_m, involves conditioning on the m-th component of the non-transformed samples, as outlined in Definition 4.1, and so unless we have that $(\boldsymbol{y}_i)_{i\in[n]} \in [1,\infty)^d$, a number of samples are necessarily lost, as highlighted in Figure 2 for the $d = 2$ case. Consequently, we are required to differentiate between the non-transformed temporal index set and the d transformed temporal index sets, which we denote by $\mathcal{I} = [n]$ and $\mathcal{I}'^{(m)} = [n^{(m)}]$ respectively. This downsampling, as well as the reduction in dimensionality, can be viewed as costs associated with obtaining multivariate Gaussian data from HR data, where we note the latter cost restricts the proposed procedure to the $d \geq 3$ case.

As discussed, the underlying precision matrices, $\Theta^{(m)}$, necessarily possess a piece-wise structure, which we further assume exhibit a degree of sparsity, as, in light of Section 2.3, this is required for meaningful conditional dependency structures within the respective GGMs. Thus, extending the aforementioned neighborhood selection procedure with an additional *fused Lasso* penalty (Tibshirani et al., 2005) is incredibly natural (Hesamian et al., 2024), and is notably the approach taken in Kolar and Xing (2012).

Specifically, for each of the sequences of random vectors, $\left(\tilde{\boldsymbol{y}}_{i'}^{(m)}\right)_{i'\in[n^{(m)}]}$, we aim to solve the following estimation procedure, for $k \in [d-1]$, and $\lambda_1, \lambda_2 \geq 0$,

$$\hat{\boldsymbol{\beta}}^{(k,m)} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{(d-2)\times n^{(m)}}}{\arg\min}\ \mathscr{L}\left(\boldsymbol{\beta}, k, m\right) + \mathscr{P}\left(\boldsymbol{\beta}, k, m, \lambda_1, \lambda_2\right), \tag{30}$$

where $\boldsymbol{\beta} = \left(\beta_{b,i'}\right)$, with $b \in [d-2]$, $i' \in [n^{(m)}]$, and

$$\mathscr{L}\left(\boldsymbol{\beta}, k, m\right) = \sum_{i'=1}^{n^{(m)}} \left(\tilde{y}_{i',k}^{(m)} - \sum_{b\in\setminus k} \tilde{y}_{i',b}^{(m)} \beta_{b,i'}\right)^2, \tag{31}$$

$$\mathscr{P}\left(\boldsymbol{\beta}, k, m, \lambda_1, \lambda_2\right) = 2\lambda_1 \sum_{i'=2}^{n^{(m)}} \|\boldsymbol{\beta}_{\cdot,i'} - \boldsymbol{\beta}_{\cdot,i'-1}\|_2 + 2\lambda_2 \sum_{i'=1}^{n^{(m)}} \sum_{b\in\setminus k} |\beta_{b,i'}|, \tag{32}$$

where an estimate, $\hat{\boldsymbol{\beta}}^{(k,m)}$, is obtained for each node, k, within the m-th $(d-1)$-dimensional GGM, and note that this node takes on the role of the response variable within the loss, $\mathscr{L}$.

Indeed, this estimation procedure, in the absence of the penalty, $\mathscr{P}$, is equivalent to the unregularised neighborhood selection procedure, whereas, when included, represents an extension of the standard neighborhood selection, which recall contains just the Lasso term (Tibshirani, 1996) in $\mathscr{P}$. Intuitively, the fused Lasso term of $\mathscr{P}$ promotes piece-wise consistency in all

components of the estimated coefficients, $\boldsymbol{\beta}$, where notably the l_2-norm is specifically utilised to encourage temporal changes to occur simultaneously across all components. Admittedly, other norms could be leveraged here, such as the l_1-norm, which is utilised within the Lasso term for its parsimonious property, however we remark its application in the fused Lasso term would result in fewer components changing, reducing the effectiveness of detecting change points that affect multiple edges simultaneously. The Lasso term of $\mathscr{P}$, on the other hand, can be understood to promote sparsity in the estimated coefficients, across both the spatial and temporal domain, and is particularly relevant in high-dimensional settings, enhancing the ability of our proposed method to detect true change points in the presence of noise. Furthermore, this sparsity, in $\boldsymbol{\beta}$, not only identifies the most relevant predictors, for the current node, k, thus improving interpretability, but also improves computational efficiency. We discuss the selection of λ_1 and λ_2 in detail within Section 5.2.3.

In a similar manner to Kolar and Xing (2012), we utilise the first differences of the aforementioned estimates, to estimate the set of partition boundaries, $\hat{\mathcal{T}}$. Specifically, we denote the matrix of first differences, for the k-th node of the m-th sequence of random vectors, by $\boldsymbol{\delta}^{(k,m)} \in \mathbb{R}^{(d-2)\times(n^{(m)}-1)}$, where matrix entries are given by

$$\delta_{b,i'}^{(k,m)} = |\hat{\beta}_{b,i'}^{(k,m)} - \hat{\beta}_{b,i'-1}^{(k,m)}|, \qquad i' \geq 2. \tag{33}$$

In light of the fused Lasso penalty term in $\mathscr{P}$, if a change point exists at the transformed temporal index, i', then we expect this to be reflected through matrix entries of $\delta_{b,i'}^{(k,m)} > 0$, for at least some of the $b \in [d-2]$, depending on the nature of the change point. Thus, it is natural to combine these first differences, at each temporal index, through a component-wise sum, so that the change point is captured irrespective of its nature. To be clear, we denote the resulting (horizontal) vector of first difference sums by $\boldsymbol{S}^{(k,m)} = \left(S_2^{(k,m)}, \ldots, S_{n^{(m)}}^{(k,m)}\right)$, where we have

$$S_{i'}^{(k,m)} = \sum_{b\in[d-2]} \delta_{b,i'}^{(k,m)} = \sum_{b\in[d-2]} |\hat{\beta}_{b,i'}^{(k,m)} - \hat{\beta}_{b,i'-1}^{(k,m)}|, \qquad i' \geq 2. \tag{34}$$

It is critical to recognise that the set of $(d-1)$ vectors of first difference sums, for each of the d sequences of random vectors, are associated with the transformed temporal index set, $\mathcal{I}'^{(m)}$. In fact, it is more illuminating to think in terms of the injective mappings, $\phi_m : \mathcal{I}'^{(m)} \mapsto \mathcal{I}$, which note are certainly known, as we have explicit knowledge of which samples were lost through conditioning. These mappings not only highlight the fact that not all elements of $\mathcal{I}$ have a corresponding preimage in $\mathcal{I}'^{(m)}$, but also that the mappings can differ, raising issues pertaining to how these vectors of first difference sums should be combined.

Of immediate concern is whether the elements of $\mathcal{T}$ have a preimage in $\mathcal{I}'^{(m)}$, for any ϕ_m, as if this is not the case, then we are prohibited from recovering change points exactly. Indeed, if we assume the probability of a sample being lost, through conditioning of the m-th auxiliary vector, is equal for all non-transformed samples, and denoted by p_m, then this occurs with probability $\prod_{m\in[d]} p_m$. Perhaps more significantly, due to the fact that different samples can be lost for

each m, even if exact change point recovery, corresponding to $\mathcal{I}'^{(m)}$, were achieved, this could potentially result in significantly more change points on $\mathcal{I}$, after applying the mappings, ϕ_m. Furthermore, the estimated change points could, in the case of subsequent missing samples, differ significantly from the true change points. This latter concern is less pressing, as we note that the expected number of samples required to observe M subsequent missing samples is given by $\frac{p_m^{-M}-1}{1-p_m}$, which in the case of $p_m = \frac{1}{2}$ reduces to $2^{M+1}-2$ samples. Indeed, not only are these issues illustrated in Figure 3, but they are also addressed in the following methodology.

First, for each $k \in [d-1]$, for each $m \in [d]$, we map the first difference sums, $S_{i'}^{(k,m)}$, from $\mathcal{I}'^{(m)}$ to $\mathcal{I}$, obtaining $S_i^{(k,m)}$, where

$$S_i^{(k,m)} = \begin{cases} S_{i'}^{(k,m)} & \text{if } \phi_m(i') = i \text{ for some } i' \in \mathcal{I}'^{(m)}, \\ 0 & \text{otherwise.} \end{cases} \tag{35}$$

Indeed, setting $S_i^{(k,m)}$ to zero reflects the absence of information due to missing samples at these indices, however it does also implicitly introduce a bias by systematically underestimating the final aggregated signal, $\dot{S}_i$, of Equation 37. This approach is conservative, as it avoids assumptions inherent in imputation methods, reducing the likelihood of false positives, and importantly ensures detected change points directly correspond to observed data.

Next, for each temporal index, we aggregate the vectors of first difference sums, which now corresponds to $\mathcal{I}$, across the $(d-1)$ nodes, resulting in $\tilde{\boldsymbol{S}}^{(m)} = \left(\tilde{S}_2^{(m)}, \ldots, \tilde{S}_n^{(m)}\right)$, where we have

$$\tilde{S}_i^{(m)} = \sum_{k \in [d-1]} S_i^{(k,m)}, \tag{36}$$

which can be considered as the output of one of d sub estimators in our proposed procedure.

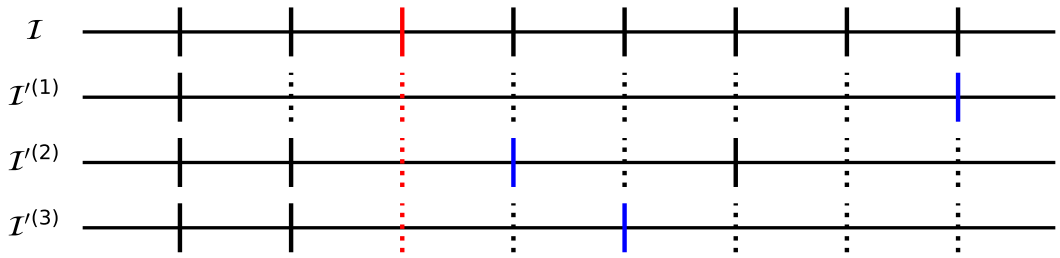

Figure 3.: The transformed temporal index sets, $\mathcal{I}'^{(m)}$, for $d = 3$, aligned with the non-transformed temporal index set, $\mathcal{I}$. Indices whose preimage is contained within $\mathcal{I}'^{(m)}$ are represented by solid lines, whereas indices corresponding to lost samples are represented by dotted lines. The true change point index, in $\mathcal{I}$, is indicated by red, whereas blue indicates optimal estimates in $\mathcal{I}'^{(m)}$, in that these are the earliest indices a change point could be detected due to lost samples.

Recalling that we began with d sequences of random vectors, we then aggregate the $\tilde{\boldsymbol{S}}^{(m)}$ across all estimation procedures, resulting in a single vector, $\dot{\boldsymbol{S}} = \left(\dot{S}_2, \ldots, \dot{S}_n\right)$, where we have

$$\dot{S}_i = \sum_{m \in [d]} \tilde{S}_i^{(m)}. \tag{37}$$

Finally, to address the aforementioned issues related to mapping between index sets, we apply smoothing across the temporal domain, via a *convolution*, utilising a kernel, $\mathcal{K}$, and subsequently estimate change points through peaks that exceed a threshold, τ. Formally, the set of estimated change points, $\hat{\mathcal{T}}$, are given by:

$$\hat{\mathcal{T}} = \left\{ i : \left(\mathcal{K} * \dot{\boldsymbol{S}}\right)_i > \tau \text{ and } \left(\mathcal{K} * \dot{\boldsymbol{S}}\right)_i > \left(\mathcal{K} * \dot{\boldsymbol{S}}\right)_{i-1}, \left(\mathcal{K} * \dot{\boldsymbol{S}}\right)_i \geq \left(\mathcal{K} * \dot{\boldsymbol{S}}\right)_{i+1} \right\}, \tag{38}$$

where the convolution operation for the one-dimensional discrete signal, $\dot{\boldsymbol{S}}$, and kernel, $\mathcal{K}$, is

$$\left(\mathcal{K} * \dot{\boldsymbol{S}}\right)_i = \sum_{l=-i}^{n-i} \dot{\boldsymbol{S}}_{i+l} \mathcal{K}(l), \tag{39}$$

noting that $\mathcal{K}$ is centered around $l = 0$.

The complete procedure is formalised in Algorithm 1, with a complete illustrated example presented in Appendix G. Although the procedure has time complexity, $\mathcal{O}\left(d^3 \cdot n_{\max} \cdot T\right)$, where $n_{\max} = \max_{m \in [d]} n^{(m)}$, and T is the total iterations required of the optimisation method utilised for this convex problem, it is notably *pleasingly parallel*, at least up until Step 9 of Algorithm 1. Indeed, Step 3 can be understood as the bottleneck of the procedure, as we expect general purpose convex optimisation algorithms to struggle, resulting in high T, although Kolar and Xing (2012) provide an accelerated gradient method of $\mathcal{O}\left(1/\epsilon\right)$, for an ϵ-accurate solution.

Algorithm 1 Proposed procedure for detecting change points in HRGMs.

Input: sequence of independent HR random vectors, $(\boldsymbol{y})_{i \in [n]} \in \mathbb{R}^d$.

Output: set of estimated change points, $\hat{\mathcal{T}} \subset \mathcal{I}$.

1: Apply φ_m to $(\boldsymbol{y})_{i \in [n]}$, for $m \in [d]$, resulting in d sets of $\left(\tilde{\boldsymbol{y}}_{i'}^{(m)}\right)_{i' \in [n^{(m)}]} \in \mathbb{R}^{d-1}$.
2: **for** $m \in [d]$ **do**
3: Obtain $(d-1)$ matrices, $\hat{\boldsymbol{\beta}}^{(k,m)} \in \mathbb{R}^{(d-2) \times n^{(m)}}$, as in Equation 30.
4: Obtain $(d-1)$ matrices, $\boldsymbol{\delta}^{(k,m)} \in \mathbb{R}^{(d-2) \times (n^{(m)}-1)}$, of first differences, as in Equation 33.
5: Obtain $(d-1)$ vectors, $\boldsymbol{S}^{(k,m)} \in \mathbb{R}^{n^{(m)}-1}$, of first difference sums, as in Equation 34.
6: Map $(d-1)$ vectors, $\boldsymbol{S}^{(k,m)}$, from $\mathcal{I}'^{(m)}$ to $\mathcal{I}$, as in Equation 35.
7: Obtain vector, $\tilde{\boldsymbol{S}}^{(m)} \in \mathbb{R}^{n-1}$, of aggregated first difference sums, as in Equation 36.
8: **end for**
9: Obtain vector, $\dot{\boldsymbol{S}} \in \mathbb{R}^{n-1}$, of aggregated sub estimators, as in Equation 37.
10: Obtain vector, $\left(\mathcal{K} * \dot{\boldsymbol{S}}\right) \in \mathbb{R}^{n-1}$, representing a smoothed version of $\dot{\boldsymbol{S}}$.
11: **for** $i \in [n-1]$ **do**
12: If $\left(\mathcal{K} * \dot{\boldsymbol{S}}\right)_i > \tau$, and a local maxima, then add i to $\hat{\mathcal{T}}$.
13: **end for**

5.2.1. Kernel Selection

The kernel, $\mathcal{K}$, encodes prior expectations regarding the temporal distribution of contributions, $\dot{\boldsymbol{S}}$, around a true change point, t_C, as well as the sensitivity and precision relative to the length of the non-transformed temporal index, n. Indeed, as we expect true change points to manifest as local maxima in $\dot{\boldsymbol{S}}$, only unimodal kernels are justified. As subsequent missing samples result in a delay of non-zero $\tilde{\boldsymbol{S}}^{(m)}$ entries, truncated left-tail, or at least right-skewed, kernels appear appropriate. However, unless we recover optimal estimates, $\boldsymbol{\beta}^{(k,m)}$, which is unlikely for non-optimal regularisation parameters, see Section 5.2.3, we expect imprecision either side of t_C, meaning a symmetric kernel may be preferable to avoid bias. Furthermore, as n increases, the likelihood of encountering M subsequent missing samples increases, whilst a given imprecision, $|\hat{t}_C - t_C|$, often becomes less problematic, and so we suggest widening $\mathcal{K}$ with increasing n to avoid overestimating the number of change points.

5.2.2. Threshold Selection

The threshold, τ, is required to differentiate between significant peaks indicative of change points, and insignificant peaks, due to noise artifacts. In this sense, the distribution of peaks, of $\left(\mathcal{K} * \dot{\boldsymbol{S}}\right)$, exhibits bimodality, where one mode corresponds to noise and the other to change points. Thus, Otsu's method (Otsu, 1979), which specifically performs an exhaustive search for the threshold that maximises the between-class variance of the two modes, provides a natural selection procedure for τ. Indeed, this method not only improves the robustness of the change point detection procedure, in that it reduces false positives, but also critically provides an automatic and adaptive thresholding mechanism applicable to various $\left(\mathcal{K} * \dot{\boldsymbol{S}}\right)$ profiles.

Importantly, in the case that $N_C = 0$, the distribution of peaks is expected to be unimodal, with this mode corresponding to noise rather than change points. In light of this, a minimum threshold, $\tau_{\min}$, is required, and ensures that the procedure remains robust in the absence of change points, which is corroborated through simulations in Section 6. If, on the other hand multimodality is expected, say for multiple change point significances, note that Otsu's method is straightforwardly extended (Reddi et al., 1984). Although, assumptions of equal variance and population of the two classes may not be met, with Kittler and Illingworth's thresholding method (Kittler and Illingworth, 1986) offering a valid alternative, we note this can be unstable in comparison to Otsu's method (Kurita et al., 1992), and thus specifically utilise the latter.

5.2.3. Regularisation Parameter Selection

The regularisation parameters, λ_1 and λ_2, control the strength of the fused Lasso and Lasso terms, of $\mathscr{P}$, respectively. Similarly to Kolar and Xing (2012), we utilise a Bayesian information criterion (BIC) type criterion for selection, due to its consistent model selection property.

Specifically, our tuning strategy is based upon the modified BIC (MBIC), as this further provides, under reasonable technical assumptions (Wang et al., 2009), model selection consistency in the case that d diverges, that being the high-dimensional setting. In the context of Equation 30, the MBIC, for node, k, of the m-th GGM, is defined as

$$\text{MBIC}^{(k,m)}\left(\lambda_1, \lambda_2\right) = \log\left(\frac{\mathscr{L}\left(\hat{\boldsymbol{\beta}}^{(k,m)}(\lambda_1, \lambda_2), k, m\right)}{n^{(m)}}\right) + \frac{\log(n^{(m)})}{n^{(m)}} \cdot \left|\left\{\delta_{b,i'}^{(k,m)} > 0\right\}\right| \cdot C_n, \quad (40)$$

where we require $C_n \to \infty$, although we note that the rate of divergence may be arbitrarily slow. In light of simulation results obtained in Wang et al. (2009), we set $C_n = \log\left(\log\left(d\right)\right)$, although remark that other choices are certainly valid.

The regularisation parameters, for the m-th GGM, are then selected such that they minimise the total MBIC across the corresponding $d-1$ nodes,

$$\left(\hat{\lambda}_1^{(m)}, \hat{\lambda}_2^{(m)}\right) = \underset{\lambda_1, \lambda_2}{\arg\min} \sum_{k \in [d-1]} \text{MBIC}^{(k,m)}\left(\lambda_1, \lambda_2\right), \quad (41)$$

where the superscript highlights that the regularisation parameters are not necessarily constant across the d GGMs. In the interest of speed and clarity, we opt to tune the regularisation parameters, as described above, over a grid of values that we explicitly specify below.

Although similar to the selection strategy of Kolar and Xing (2012), this is importantly the most problematic aspect of our proposed procedure. From a computational perspective, the bottleneck of our procedure must be run for each parameter pair we wish to evaluate, and, in the absence of resource to leverage the pleasingly parallel nature of Algorithm 1, this can thus be prohibitively slow, particularly for large d. Furthermore, the estimated change points, $\hat{\mathcal{T}}$, are often sensitive to the $\hat{\lambda}_1^{(m)}$ and $\hat{\lambda}_2^{(m)}$ values, suggesting a high resolution grid is required. Lastly, note that in the case of $\hat{\lambda}_1^{(m)}, \hat{\lambda}_2^{(m)} = 0$, the estimates, $\hat{\boldsymbol{\beta}}^{(k,m)}$, minimise $\mathscr{L}$, as $\mathscr{P} = 0$, and since we empirically observe the former term in Equation 40 to dominate the latter, we can expect this parameter pair to be selected, if present in the grid.

Thus, we propose to tune over a grid bounded closely above and below values of λ_1 and λ_2 that provide convergence rate guarantees, at least in the absence of the other regularisation parameter, for the estimates, $\hat{\boldsymbol{\beta}}^{(k,m)}$, in relation to the true solution $\boldsymbol{\beta}_*^{(k,m)}$. Regarding λ_2, when $\lambda_1 = 0$, Equation 30 reduces to the standard Lasso case, which is shown under reasonable technical conditions (Wainwright, 2009) to have convergence rate,

$$\|\hat{\boldsymbol{\beta}}^{(k,m)} - \boldsymbol{\beta}_*^{(k,m)}\|_2 = \mathcal{O}\left(\sqrt{\frac{l(d-1)}{n^{(m)}}}\right), \quad \text{if } \lambda_2 = \Theta\left(\sqrt{\frac{\log(d-1)}{n^{(m)}}}\right), \quad (42)$$

where l is the number of zero indices in $\boldsymbol{\beta}_*^{(k,m)}$, and $a_n = \Theta(b_n)$ denotes the sequence a_n being asymptotically bounded, both above and below, by the sequence b_n, rather than just above

as indicated by the standard notation, $\mathcal{O}(a_n)$. In fact, in corroboration with this author, in Ravikumar et al. (2010) and Hastie et al. (2015), we suggest utilising $\check{\lambda}_2^{(m)} = 2\sqrt{\frac{\log(d-1)}{n^{(m)}}}$.

On the other hand, if $\lambda_2 = 0$, then Equation 30 reduces to the 1-dimensional fused Lasso scenario, which is notably a special case of the general fused Lasso convex optimisation problem (Tibshirani and Taylor, 2011), in which the graph associated with first differences is a *chain graph*. As the number and location of true change points are unknown, the minimum difference between indices at which non-zero first differences occur in $\boldsymbol{\beta}_*^{(k,m)}$ can feasibly be $n^{(m)}$, meaning that the MSE convergence rate outlined in Padilla et al. (2018) becomes

$$\|\hat{\boldsymbol{\beta}}^{(k,m)} - \boldsymbol{\beta}_*^{(k,m)}\|_{n^{(m)}}^2 = \mathcal{O}\left(\frac{l}{n^{(m)}}\left(\log(l) + \log(\log(n^{(m)}))\log(n^{(m)}) + 1\right)\right), \tag{43}$$

provided $\lambda_1 = \Theta\left(\sqrt{n^{(m)}}\right)$. In fact, when $|\mathcal{T}| = \mathcal{O}(1)$, Theorem 2 of Lin et al. (2016) becomes particularly useful, and suggests that λ_1 should be of approximately order $\sqrt{n^{(m)}}$, as does Theorem 4 of the respective paper. In light of simulations in Section 6, we suggest that utilising $\check{\lambda}_1^{(m)} = \sqrt{n^{(m)}}$ provides reasonable results.

5.2.4. Potential Extensions

It is not unreasonable, in many real-world applications, to expect that additional information pertaining to the change points exists, as we may have, due to domain expertise, prior beliefs, or even explicit knowledge, regarding the number or location of change points. In the following, we focus on the former scenario, as it is intuitive to extend our proposed procedure, however, we note that the latter scenario can be addressed by enforcing sparsity in the $\hat{\boldsymbol{\beta}}^{(k,m)}$ estimates within temporal partitions in which we do not expect change points to occur.

Indeed, consider the case in which we have a prior probabilistic belief, $p(N_C)$, regarding the number of change points, N_C, and note that it is natural to augment the MBIC such that goodness-of-fit reflects how well the $\boldsymbol{S}^{(k,m)}$, and therefore $\hat{N}_C$, aligns with $p(N_C)$, where specifically we suggest,

$$\text{MBIC}_{\text{aug}}^{(k,m)}(\lambda_1, \lambda_2) = \text{MBIC}^{(k,m)}(\lambda_1, \lambda_2) - 2\log\left(p\left(\left|\left\{S_{i'}^{(k,m)} > 0\right\}\right|\right)\right), \tag{44}$$

which, in the case that we assume $N_C \sim \mathcal{N}(\mu, \sigma^2)$, effectively reduces to

$$\text{MBIC}_{\text{aug}}^{(k,m)}(\lambda_1, \lambda_2) = \text{MBIC}^{(k,m)}(\lambda_1, \lambda_2) + \frac{\left(\left|\left\{S_{i'}^{(k,m)} > 0\right\}\right| - \mu\right)^2}{\sigma^2}. \tag{45}$$

On the other hand, if we have explicit knowledge regarding a minimum, maximum, or exact number of change points, then a compatible number of peaks in decreasing magnitude, rather than all peaks, above τ could instead be selected. In any case, a significant advantage of the proposed procedure is its potential adaptability to different use cases.

5.3. Optimal Estimates

Motivated by the fact that Equation 30 is convex, rather than strictly convex, implying multiple solutions exist, we provide, as in Kolar and Xing (2012), conditions characterising optimal estimates of $\hat{\boldsymbol{\beta}}^{(k,m)}$ in the unrestricted case, where the number and location of change points are unknown, as is the sparsity pattern of the first differences of the true $\boldsymbol{\beta}^{(k,m)}$.

Lemma 5.1. The matrix, $\hat{\boldsymbol{\beta}}^{(k,m)}$, is optimal for the convex optimisation problem of Equation 30 if and only if there exists a collection of sub-gradient vectors, $\{\hat{\boldsymbol{z}}_{i'}\}_{i' \in [2:n^{(m)}]}$ and $\{\hat{\boldsymbol{w}}_{i'}\}_{i' \in [n^{(m)}]}$, with $\hat{\boldsymbol{z}}_{i'} \in \partial \|\hat{\boldsymbol{\beta}}^{(k,m)}_{\cdot,i'} - \hat{\boldsymbol{\beta}}^{(k,m)}_{\cdot,i'-1}\|_2$ and $\hat{\boldsymbol{w}}_{i'} \in \partial \|\hat{\boldsymbol{\beta}}^{(k,m)}_{\cdot,i'}\|_1$, satisfying, for all $l' \in [n^{(m)}]$,

$$\mathbf{0} = \sum_{i'=l'}^{n^{(m)}} (-\tilde{\boldsymbol{y}}^{(m)}_{i',\setminus k}) \left(\tilde{y}^{(m)}_{i',k} - \sum_{b \in \setminus k} \tilde{y}^{(m)}_{i',b} \hat{\beta}^{(k,m)}_{b,i'} \right) + \lambda_1 \hat{\boldsymbol{z}}_{l'} + \lambda_2 \sum_{i'=l'}^{n^{(m)}} \hat{\boldsymbol{w}}_{i'}, \tag{46}$$

Proof. See Appendix H.1. □

In fact, as outlined in the Proof of Lemma 5.1, Equation 46 can be written solely in terms of an estimate of the matrix, $\boldsymbol{\gamma}$, which we introduce as

$$\boldsymbol{\gamma}^{(k,m)}_{i'} = \begin{cases} \boldsymbol{\beta}^{(k,m)}_{\cdot,i'}, & \text{for } i' = 1, \\ \boldsymbol{\beta}^{(k,m)}_{\cdot,i'} - \boldsymbol{\beta}^{(k,m)}_{\cdot,i'-1}, & \text{otherwise.} \end{cases} \tag{47}$$

Notably, we further derive the optimally conditions in the restricted case, in which the number and location of change points are explicitly known, as is the sparsity pattern of the first differences of the true $\boldsymbol{\beta}^{(k,m)}$.

Lemma 5.2. If the number, N_C, and locations, $\mathcal{T} = \{t_1, \ldots, t_{N_C}\}$, of change points are explicitly known, and $\boldsymbol{\beta}^{(k,m)}$ is known to be constant within resulting partitions, then the matrix, $\hat{\boldsymbol{\beta}}^{(k,m)}$, is optimal for the corresponding restricted convex optimisation problem of Equation 30 if and only if $\hat{\boldsymbol{\gamma}}^{(k,m)}$ satisfies a collection of $(N_C + 1)$ equations, as characterised in Equation 69.

Proof. See Appendix H.2 □

Although not yet explicitly derived, values of $\boldsymbol{\gamma}$ satisfying both the set of $n^{(m)}$ optimality equations, corresponding to the unrestricted case, and the $(N_C + 1)$ optimality equations, corresponding to the restricted case, then form a set of estimates we can consider. Specifically, these estimates are guaranteed to recover the correct number and locations of change points, at least on the transformed temporal index set, along with the correct sparsity pattern, and is thus an interesting problem left for future research.

6. Simulation Studies

To illustrate the effectiveness of the procedure outlined in Section 5.2, we present a suite of simulation studies that aim to provide comprehensive coverage over various (N_C, n, d) triplets. In contrast to Kolar and Xing (2012), our procedure estimates only change point locations, resulting in $(\hat{N}_C+1)$ disjoint partitions of $\mathcal{I}$, denoted by P_α, such that $\bigcup_{\alpha=1}^{\hat{N}_C+1} P_\alpha = [n]$. Indeed, it is the structure of the corresponding HRGMs, rather than the GGMs, that are ultimately of interest, and, as in Section 7, we suggest recovering these structures through the `EGlearn` algorithm of Engelke et al. (2024). Thus, it is natural to evaluate $\hat{\mathcal{T}}$ using clustering metrics, where, specifically, we utilise the adjusted Rand index (ARI) (Hubert and Arabie, 1985), as it provides an interpretable yet robust measure of agreement between these partitions (Santos and Embrechts, 2009), and is the evaluation metric of choice in Milligan and Cooper (1986).

Indeed, the ARI is a *corrected-for-chance* extension of the Rand index (RI) (Rand, 1971), where the latter metric is defined as the accuracy of all assignment decisions,

$$\text{RI} = \frac{a+b}{a+b+c+d} = \frac{a+b}{{}^{n}C_2},$$

where a denotes the number of pairs, $(i,j) \in \mathcal{I}$, that belong to the same partition, whilst b denotes the number of pairs that belong to different partitions, across $\hat{\mathcal{T}}$ and $\mathcal{T}$, whereas c and d denote the cases of disagreement across the sets of change points. Importantly, for two sets of random partitions, the expected RI is neither constant, nor necessarily zero, complicating interpretation of RI values across different simulations, and thus motivates the use of the ARI,

$$\text{ARI} = \frac{\text{RI} - \mathbb{E}\{\text{RI}\}}{\max(\text{RI}) - \mathbb{E}\{\text{RI}\}} = \frac{\text{RI} - \mathbb{E}\{\text{RI}\}}{1 - \mathbb{E}\{\text{RI}\}},$$

which is also bounded above by one, but notably has an expected value of zero when partitions are random. Whilst the ARI provides an overall measure of performance, it is adversely affected by both misestimation of N_C and imprecise estimates. Thus, to disentangle these effects, and to allow for more nuanced interpretation, we additionally report $\Delta N_C = |\hat{N}_C - N_C|$ and ARI^* values, where the latter metric is the ARI conditioned on $\hat{N}_C = N_C$.

In our simulations, data is generated in a deterministic and reproducible manner, meaning that the term *random*, as used below, refers to psuedo-randomness, rather than true randomness. First, a sequence of graphs, $(\mathcal{G}_\alpha)_{\alpha\in[N_C+1]}$, associated to the underlying HRGMs, are generated in a similar manner to Kolar et al. (2010), where given an initial sparsity level, ρ_0, we randomly generate, as described in Erdos and Renyi (1960), a connected graph, $\mathcal{G}_0$, with $|E_0| = \lfloor \rho_0 \cdot {}^{d}C_2 \rfloor$ edges. Then, given a change sensitivity index, κ, and a minimum number of edge changes, Δ_{min}, we calculate the number of edges to change, $\delta_\alpha = \max(\lfloor \kappa \cdot |E_\alpha| \rfloor, \Delta_{min})$, where edges are added and removed with equal probability, to obtain $\mathcal{G}_\alpha$, for $\alpha \in [N_C+1] \setminus \{1\}$. We ensure that graphs remain connected during this process, and refer to Figure 7 for an illustration of the process. Then, for each $\mathcal{G}_\alpha$, a valid random extremal variogram matrix, Γ_α, is generated and associated, provided that it is a distinct matrix. Indeed, identical graphs are permitted,

and are certainly necessary when $N_C \geq 2^{d C_2}$, which is the bound when disconnected graphs are also considered.

Next, the temporal index set of true change points, $\mathcal{T}$, are randomly sampled without replacement, such that change point indices, $\{t_\alpha\}_{\alpha \in [N_C+1]}$, are never less than $dt_{\min}$ indices from each other. To mitigate systematic bias potentially introduced in the data generation process, we choose a random permutation of the $(\Gamma_\alpha)_{\alpha \in [N_C+1]}$ sequence, in that we thus expect to capture cases where the edge sets, E_α and $E_{\alpha+1}$, are equal before and after t_α, as well as cases in which they differ. Finally, utilising the respective extremal variogram, Γ_α, for the current partition, P_α, we generate n exact HR samples, to which we apply Algorithm 1.

Notably, this generation procedure reflects how we expect subtle, and thus difficult to detect, extremal conditional dependency changes to occur in practise, as, for example, it allows for an initial small change in the HRGM edge set, followed by the reversal in the subsequent change point, resulting in the original edge set. Intuitively, the data generating hyper parameters, κ and $\Delta_{\min}$, control the severity of change points, whereas $dt_{\min}$ establishes a lower bound for the separation between change point locations. In the interest of difficulty, we run our simulations for relatively low values of $\kappa = 0.3$, $\Delta_{\min} = 1$, and $dt_{\min} = \lfloor \frac{n}{40} \rfloor$. Furthermore, we set $\rho_0 = 0.1$, and evaluate our procedure, for each (N_C, n, d) triplet, over 50 simulations.

Regarding the hyper parameters of Algorithm 1, we adopt a 1D Gaussian kernel with standard deviation, $\sigma = \frac{n}{100}$, aligning with the rationale discussed in Section 5.2.1. Furthermore, the threshold is selected as discussed in Section 5.2.2, with $\tau_{\min} = \frac{d(d-1)}{4000}$, reflecting the noise level we consider negligible in the first difference sums, $\boldsymbol{S}^{(k,m)}$. Indeed, $\lambda_1^{(m)}$ and $\lambda_2^{(m)}$ are tuned as outlined in Section 5.2.3, over 20 equally spaced values, between $[\frac{1}{3}\check{\lambda}_1^{(m)}, 3\check{\lambda}_1^{(m)}]$, and $[\frac{1}{3}\check{\lambda}_2^{(m)}, 3\check{\lambda}_2^{(m)}]$ respectively. Although a finer grid would ideally be utilised, we argue this is a constraint consistent with resource-limited applications, particularly in cases where d is significantly large.

Results obtained for $N_C = 1$ are reported below, in Table 1, and we refer to Appendix F for all results, where the $N_C = 0, 1, 2$ cases are reported in Tables 5, 6, and 7 respectively.

d \ n	250	500	1000	2000
3	(0.32, 0.694, 0.950)	(0.22, 0.801, 0.947)	(0.26, 0.729, 0.975)	(0.28, 0.715, 0.993)
4	(0.48, 0.772, 0.948)	(0.16, 0.908, 0.963)	(0.06, 0.955, 0.984)	(0.20, 0.861, 0.963)
5	(0.14, 0.923, 0.966)	(0.08, 0.939, 0.977)	(0.06, 0.953, 0.974)	(0.06, 0.961, 0.991)
7	(0.40, 0.909, 0.970)	(0.18, 0.920, 0.975)	(0.10, 0.940, 0.988)	(0.24, 0.934, 0.990)
10	(0.16, 0.946, 0.982)	(0.18, 0.917, 0.980)	(0.16, 0.950, 0.987)	(0.18, 0.929, 0.992)
15	(0.52, 0.874, 0.964)	(0.10, 0.951, 0.973)	(0.06, 0.941, 0.960)	(0.06, 0.963, 0.985)
20	(0.48, 0.851, 0.939)	(0.26, 0.914, 0.966)	(0.00, 0.986, 0.986)	(0.04, 0.959, 0.982)

Table 1.: Average values of $(\Delta N_C, \text{ARI}, \text{ARI}^*)$, for $N_C = 1$ simulation study triplets, obtained over 50 simulations.

Although the data generating procedure facilitates an assessment regarding the robustness of our proposed procedure, the pseudo-randomness employed results in change point detection problems of varying difficulty, and, even when averaging over 50 simulations, we expect variability in performance that can be attributed to characteristics of the generated data, rather than the simulation study triplet values. Indeed, we expect simulations in which change points are close to each other, or near the boundaries of $\mathcal{I}$, that is $i = 1$ and $i = n$, to be of particular challenge, and further expect cases where $E_\alpha = E_{\alpha+1}$ to heighten the difficulty. Thus, specific comparison of (N_C, n, d) triplets should be approached with caution, and instead we focus on general trends across the triplet parameters.

Of particular interest is the $N_C = 0$ case, as this explicitly informs the global false positive rate, or equivalently the global specificity, of our proposed procedure, thus determining the need of an additional screening test for the presence of change points. In general, a remarkable specificity is observed, alleviating the need for such a test, although for smaller sample sizes, and increasing dimensionality, we do observe an increased propensity for false positives, which aligns with the curse of dimensionality. Despite our specification of $\tau_{\min}$, in high-dimensional spaces it becomes increasingly difficult to discern true structural changes, due to the fact that the total number of pairwise relationships grows quadratically in d, increasing the likelihood and magnitude of noise. Although mitigated by increasing sample size, such increases are not always feasible, and so we propose further refining the regularisation selection procedure, as well as the specification of $\tau_{\min}$, in such scenarios.

Importantly, we observe our procedure to maintain robustness in the presence of true change points, with generally large ARI scores, and particularly high ARI* scores, where the latter indicates that misestimation of N_C, rather than estimated change point location, is problematic for our procedure. Upon investigation, we observe the procedure to suffer from over-estimation only in high dimensions, again highlighting issues of false positives in these scenarios, whereas in lower dimensions it suffers from both under-estimation and over-estimation. In general, we observe an increase in d to result in higher ARI scores, provided n is large enough to mitigate the aforementioned curse of dimensionality, and can intuitively be explained through Condorcet's Jury Theorem (Ladha, 1995), as our procedure can in fact be viewed as an aggregation of d sub estimators. Similarly, we observe increasing n to result in larger ARI scores, which further aligns with the convergence rates of Equations (42) and (43). Notably, for the $d = 3$ case, misestimation is extremely prevalent, as reflected through the significantly lower ARI scores for all values of n. Indeed, this represents the minimal non-trivial case for our procedure, in which aggregation, for each node, is performed on only two vectors, rather than matrices, significantly restricting the available information leveraged for change point detection. In this sense, the benefits of our multi-dimensional approach are not utilised, and the nuanced balance between dimensionality and effective change point detection is again highlighted.

As expected, degradation in the evaluation metrics are generally observed for increasing N_C, as the estimation procedure becomes more challenging, since not only are there more change points, but also a higher likelihood of change points occurring close to each other, or near

the boundaries of $\mathcal{I}$. Furthermore, for $N_C > 1$, we note that the worst-case permutation of $(\Gamma_\alpha)_{\alpha\in[N_C+1]}$, in which $E_\alpha = E_{\alpha+1}$, can potentially be obtained for any simulation. The reduction in performance, however, appears to be less pronounced in the regime of large n, suggesting increasing n also helps to offset issues associated with increasing N_C. In any case, the results of the simulation study suggest that extending the procedure, as discussed in Section 5.2.4, is a promising direction for lower values of n, particularly if used to iteratively detect the most prominent change point, until no change points are detected in the resulting partitions. We conclude by noting that although of great value, extending the simulation studies across more triplet values is out of the scope of the paper due to available resource, as is a comprehensive sensitivity analysis for the data generating hyper parameters.

7. Applications

Unlike change point detection methodologies for GGMs, and even the simulation studies considered in Section 6, a critical distinction of real world applications is that we observe data, $\boldsymbol{X}_i$, that is considered to be in the Pareto max-domain of attraction of the MPD, $\boldsymbol{Y}_i$, rather than $\boldsymbol{Y}_i$ samples directly. Noting that $\{F_m(X_{i,m}) > 1-q\} \subset \{F(\boldsymbol{X}) \not\leq 1-q\}$, we have that if the limit in Equation (18) holds, then as $q \to 0$, we necessarily have

$$\frac{q}{1-F(\boldsymbol{X}_i)} \,\middle|\, \{F_m(X_{i,m}) > 1-q\} \xrightarrow{\mathrm{D}} \boldsymbol{Y}_i^{(m)}, \tag{48}$$

where notably convergence is to $\boldsymbol{Y}_i^{(m)}$, rather than $\boldsymbol{Y}_i$, for which we refer to Proposition S4 of Engelke et al. (2024) for the technical details in the interest of brevity. Thus, as outlined in the aforementioned paper, if we consider the empirical $d.f.$ of $\boldsymbol{X}_i$, which to be clear is given by $\tilde{F}(\boldsymbol{x}) = \left(\tilde{F}_1(x_1), \ldots, \tilde{F}_d(x_d)\right)$, where $\tilde{F}_m(x_m) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_{i,m} \leq x_m\}$, then provided $\frac{k}{n}$ is sufficiently small, we crucially observe

$$\hat{\boldsymbol{Y}}_i^{(m)} = \left\{ \frac{k}{n(1-\tilde{F}(\boldsymbol{X}_i))} : \tilde{F}_m(X_{i,m}) > 1 - \frac{k}{n} \right\} \tag{49}$$

to form an approximate sample from $\boldsymbol{Y}_i^{(m)}$, which contains the $k \in [n-1]$ most extreme observations of component m, since $\tilde{F}_m$ effectively ranks observations in ascending order. In fact, the selection of k is a persisting challenge in the extremal setting and amounts to a bias-variance trade-off (Engelke and Volgushev, 2022), for which we suggest addressing through a stability analysis of the estimated extremal correlation, $\hat{\chi}_{\alpha\beta}$, in a similar manner to the Hill method (Drees et al., 2000), which we notably provide details for in Appendix D. Indeed, $\hat{\chi}_{\alpha\beta}$ is the empirical estimate of the extremal correlation, $\chi_{\alpha\beta}$, a popular summary statistic for the extremal conditional dependency structure of the bivariate margins for $\alpha, \beta \in [d]$, and is defined through

$$\chi_{\alpha\beta} = \lim_{q\to 0} \mathbb{P}\left(F_\alpha(X_{i,\alpha}) > 1-q \mid F_\beta(X_{i,\beta}) > 1-q\right) \in [0,1], \tag{50}$$

meaning that $\hat{\chi}_{\alpha\beta}$ is defined as

$$\hat{\chi}_{\alpha\beta} = \frac{n}{k} \sum_{i=1}^{n} \mathbb{1} \left\{ \tilde{F}_{\alpha}(X_{i,\alpha}) > 1 - \frac{k}{n},\ \tilde{F}_{\beta}(X_{i,\beta}) > 1 - \frac{k}{n} \right\}. \tag{51}$$

To illustrate an application, and the value, of our proposed procedure, we consider $n = 3790$ daily spot foreign exchange rates, expressed in terms of British pound sterling, of $d = 26$ currencies, from 1 October 2005 to 30 September 2020, as introduced in Engelke and Volgushev (2022) and also considered in Engelke et al. (2024). Indeed, careful consideration regarding the validity of the underlying model assumptions is required, and in particular, as financial time series often exhibit significant temporal dependence structure, pre-processing is necessary to obtain approximately independent samples. To this end, as outlined in Engelke and Volgushev (2022), an ARMA(0,2)-GARCH(1,1) model, with estimated mean, $\hat{\mu}_{i,m}$, and standard deviation, $\hat{\sigma}_{i,m}$, is utilised to filter the daily log-returns, $R_{i,m}$, where $i \in [n]$, $m \in [d]$. In particular, we are interested in extremes in both directions, and thus the absolute values of the standardised filtered returns are considered,

$$X_{i,m} = \left| \frac{R_{i,m} - \hat{\mu}_{i,m}}{\hat{\sigma}_{i,m}} \right|. \tag{52}$$

Unlike Engelke et al. (2024), in the interest of utilising a finer grid of regularisation parameters in our tuning procedure, and to enhance clarity regarding structural changes, we take the perspective of developing tail risk hedging strategies for a selected subset of $d = 10$ exchange rates, namely those present in Figure 4. To be clear, given a selected value of k, for each component of $(\boldsymbol{X}_i)_{i\in[n]}$, observations are empirically ranked and transformed to a standard Pareto scale, then filtered such that only observations corresponding to one of the k most extreme exceedances of a component are kept, and lastly standardised to a MPD scale. Through inspection of various stability plots in Figure 6, we select $k = \lfloor n^{0.75} \rfloor = 483$, which notably satisfies the assumptions of Theorem 1 in Engelke et al. (2024), and is slightly higher than in this paper's corresponding application, which in light of the impact of the effective sample size observed in Section 6 appears advantageous for our proposed procedure, particularly given the feasibility of encountering multiple change points.

For consistency, we adopt identical hyper parameter values for Algorithm 1 as presented in Section 6, although now notably tune over 40, rather than 20, equally spaced values in the respective regularisation parameter ranges. Importantly, note that $\hat{\boldsymbol{Y}}_i^{(m)}$ now serves as the input to Algorithm 1, and since observations are lost through obtaining $\hat{\boldsymbol{Y}}_i^{(m)}$, we must in fact map $\hat{\mathcal{T}}$ back to $\mathcal{I} = [n]$, analogous to how ϕ_m mapped $\mathcal{I}'^{(m)}$ to $\mathcal{I}$. Accounting for this, we obtain $\hat{\mathcal{T}} = \{837, 1128, 1497, 2099\}$, and subsequently recover estimates of HRGMs, for the resulting partitions, through the `EGlearn` algorithm, following the approach of Engelke et al. (2024), and present these structures in Figure 4. Remarkably, the first estimated change point coincides with the conclusion of the global financial crisis of 2007-2008, which profoundly impacted

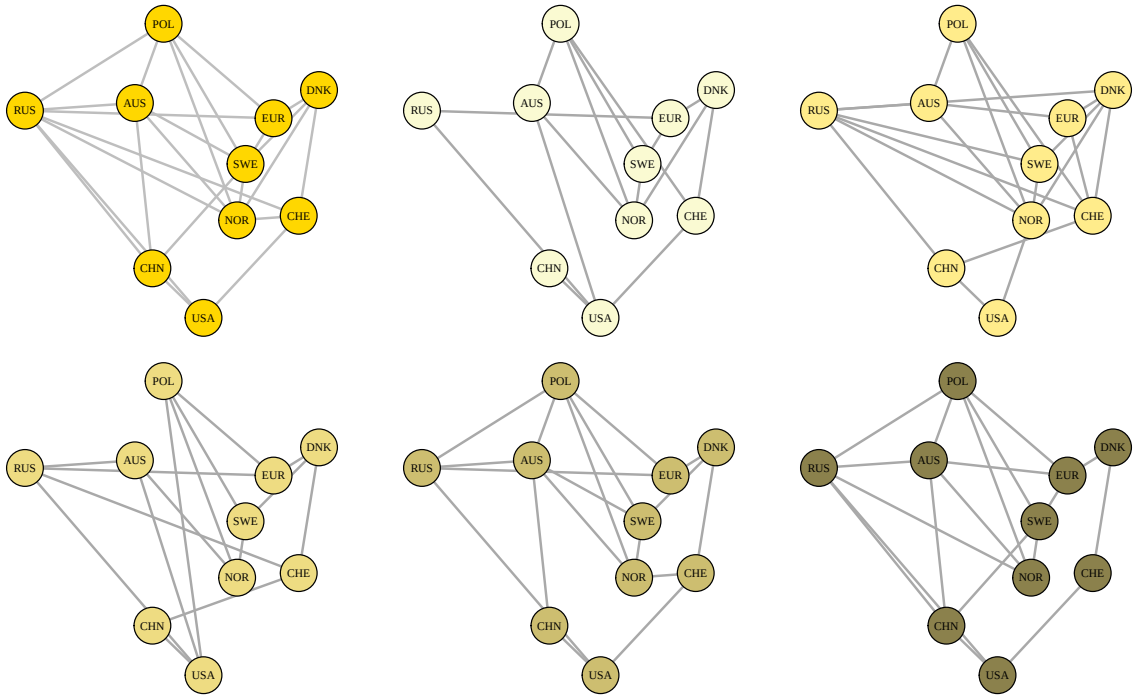

Figure 4.: The HRGMs estimated by `EGlearn`, and selected by MBIC, for a subset of $d = 10$ currencies, for the entirety of the exchange rates dataset (top-left), and for the partitions resulting from $\hat{\mathcal{T}}$ (top-middle to bottom-right, from light to dark). The list of currencies with the corresponding abbreviations are shown in Table 3.

global exchange rates (Melvin and Taylor, 2009), and notably we observe a brief increase in the extremal conditional dependencies involving the Swiss Franc and other currencies following this change point, which is consistent with our expectation that such a safe haven currency would become central to the HRGMs during times of uncertainty. Additionally, the third estimated change point corresponds exactly to the Swiss National Bank's bid to peg the Swiss Franc against the Euro (Hui et al., 2016), whereas the fourth estimated change point is particularly close to the first round of Western sanctions against Russia in 2014, which have been theorised to predicate the subsequent collapse of the Russian Ruble (Viktorov and Abramov, 2020).

8. Discussion

This paper introduces a novel, yet straightforward and intuitive approach for offline change point detection in the extremal setting, which naturally complements existing methodologies concerning the estimation of HRGMs. In light of the simulations considered in Section 6, the procedure effectively serves the dual purpose of screening and estimation, demonstrating promising performance in the case of exact HR samples, for which true change points were purposely subtle. Remarkably, the effect of increasing dimension is observed to be generally beneficial to the effectiveness of our procedure, or at least not noticeably detrimental, provided a large enough effective sample size. Indeed, establishing a theoretical justification for

this phenomenon would be of notable value, and also overcome current limitations related to interpretation due to the significant number of hyper parameters involved in both the change point detection and data generating procedures.

Despite this, due to limitations in the resource and efficiency of our implementation, simulations were admittedly restricted to moderate dimensions and a relatively low number of true underlying change points, as was the resolution of the regularisation parameter grid we tuned over constrained. Indeed, the time complexity, and strategy utilised to tune the regularisation parameters, are the most notable shortcomings of our proposed procedure, particularly in the absence of ideal constants to bound λ_1 and λ_2. To this end, enhancing the implementation of Algorithm 1 and refining the tuning methodology thus present valuable *low-hanging fruit*.

The utility, and applicability, of our proposed procedure is aptly demonstrated in context of foreign exchange rates, where multiple HRGMs exhibiting greater sparsity, and revealing dependency changes, are recovered, that are otherwise obfuscated when assuming a single static structure. Whilst estimated change points align with global events expected to shock currency markets, the notable absence of the onset of the global financial crisis of 2007-2008 underscores a common difficulty in change point detection, namely that of accurate estimation near temporal index set boundaries. Although this application is a single illustrative example, chosen to highlight synergy with the `EGlearn` algorithm, our procedure is widely applicable, without domain expertise, to almost any multivariate extremal problem (Engelke and Hitz, 2020), provided the assumption of a HRGM is reasonable.

Indeed this paper primarily lays the foundational groundwork upon which more complex and domain-specific change point detection methodologies, such as the extremal counterpart of Londschien et al. (2021), can be developed. Perhaps more importantly, however, is the explicit demonstration of how the transformation outlined in Lemma 3.1 facilitates the employment of Gaussian techniques, in an aggregational manner, within the extremal setting. Notably, instead of estimating the extremal variogram directly, one could, for example, reasonably utilise the majority vote from d precision matrix estimates obtained from either one of the aforementioned base learners, or even a more sophisticated procedure, such as in Wang et al. (2016).

References

Abad, P., Benito, S., and López, C. (2014). A comprehensive review of value at risk methodologies. *The Spanish Review of Financial Economics*, 12(1):15–32.

Bücher, A. and Zhou, C. (2021). A Horse Race between the Block Maxima Method and the Peak–over–Threshold Approach. *Statistical Science*, 36(3):360 – 378.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. Springer-Verlag, London.

Cooley, D., Davis, R. A., and Naveau, P. (2010). The pairwise beta distribution: A flexible parametric multivariate model for extremes. *Journal of Multivariate Analysis*, 101(9):2103–2117.

de Carvalho, M. and Davison, A. C. (2014). Spectral density ratio models for multivariate extremes. *Journal of the American Statistical Association*, 109(506):764–776.

de Fondeville, R. and Davison, A. C. (2018). High-dimensional peaks-over-threshold inference. *Biometrika*, 105(3):575–592.

Drees, H., Resnick, S., and de Haan, L. (2000). How to make a Hill plot. *The Annals of Statistics*, 28(1):254–274.

Engelke, S. and Hitz, A. S. (2020). Graphical models for extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):871–932.

Engelke, S. and Ivanovs, J. (2021). Sparse structures for multivariate extremes. *Annual Review of Statistics and Its Application*, 8(Volume 8, 2021):241–270.

Engelke, S., Lalancette, M., and Volgushev, S. (2024). Learning extremal graphical structures in high dimensions.

Engelke, S. and Taeb, A. (2024). Extremal graphical modeling with latent variables.

Engelke, S. and Volgushev, S. (2022). Structure Learning for Extremal Tree Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):2055–2087.

Erdos, P. and Renyi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hungary. Acad. Sci.*, 5:17–61.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2):180–190.

Fougères, A.-L. (2003). Multivariate extremes. *Extreme Values in Finance, Telecommunications, and the Environment*, pages 373–388.

Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Green, P. J. and Thomas, A. (2013). Sampling decomposable graphs using a Markov chain on junction trees. *Biometrika*, 100(1):91–110.

Gudendorf, G. and Segers, J. (2010). Extreme-value copulas. In Jaworski, P., Durante, F., Härdle, W. K., and Rychlik, T., editors, *Copula Theory and Its Applications*, pages 127–145, Berlin, Heidelberg. Springer Berlin Heidelberg.

Gumbel, E. J. (1958). *Statistics of Extremes*. Columbia University Press, New York Chichester, West Sussex.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC.

Hesamian, G., Johannssen, A., and Chukhrova, N. (2024). An explainable fused lasso regression model for handling high-dimensional fuzzy data. *Journal of Computational and Applied Mathematics*, 441(C).

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Hui, C.-H., Lo, C.-F., and Fong, T. P.-W. (2016). Swiss franc's one-sided target zone during 2011–2015. *International Review of Economics & Finance*, 44:54–67.

Hüsler, J. and Reiss, R.-D. (1989). Maxima of normal random vectors: Between independence and complete dependence. *Statistics & Probability Letters*, 7(4):283–286.

Jordan, M. I. (2004). Graphical Models. *Statistical Science*, 19(1):140–155.

Kittler, J. and Illingworth, J. (1986). Minimum error thresholding. *Pattern Recognition*, 19(1):41–47.

Kolar, M., Song, L., Ahmed, A., and Xing, E. P. (2010). Estimating time-varying networks. *The Annals of Applied Statistics*, 4(1):94–123.

Kolar, M. and Xing, E. P. (2012). Estimating networks with jumps. *Electronic Journal of Statistics*, 6(none):2069–2106.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press.

Kotz, S. and Nadarajah, S. (2000). *Extreme value distributions : theory and applications*. Imperial College Press, London.

Kurita, T., Otsu, N., and Abdelmalek, N. (1992). Maximum likelihood thresholding based on population mixture models. *Pattern Recognition*, 25(10):1231–1240.

Ladha, K. K. (1995). Information pooling through majority-rule voting: Condorcet's jury theorem with correlated votes. *Journal of Economic Behavior & Organization*, 26(3):353–372.

Lalancette, M. (2023). On pairwise interaction multivariate pareto models. *Stat*, 12(1):e613.

Lauritzen, S. (1996). *Graphical Models*. Oxford Statistical Science Series. Clarendon Press.

Lin, K., Sharpnack, J., Rinaldo, A., and Tibshirani, R. J. (2016). Approximate recovery in changepoint problems, from ℓ_2 estimation error rates.

Londschien, M., Kovács, S., and Bühlmann, P. (2021). Change-point detection for graphical models in the presence of missing values. *Journal of Computational and Graphical Statistics*, 30(3):768–779.

Maathuis, M., Drton, M., Lauritzen, S., and Wainwright, M., editors (2019). *Handbook of Graphical Models*. Handbooks of Modern Statistical Methods. CRC Press.

Manuel Hentschel, S. E. and Segers, J. (2024). Statistical inference for hüsler–reiss graphical models

through matrix completions. *Journal of the American Statistical Association*. To appear.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462.

Melvin, M. and Taylor, M. P. (2009). The crisis in the foreign exchange market. *Journal of International Money and Finance*, 28(8):1317–1330. The Global Financial Crisis: Causes, Threats and Opportunities.

Milligan, G. W. and Cooper, M. C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, 21(4):441–458. PMID: 26828221.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66.

Padilla, O. H. M., Sharpnack, J., Scott, J. G., and Tibshirani, R. J. (2018). The dfs fused lasso: Linear-time denoising over general graphs. *Journal of Machine Learning Research*, 18(176):1–36.

Papastathopoulos, I. and Strokorb, K. (2016). Conditional independence among max-stable laws. *Statistics & Probability Letters*, 108:9–15.

Proschan, M. and Shaw, P. (2016). *Essentials of Probability Theory for Statisticians*. A Chapman & Hall book. CRC Press, Taylor & Francis Group.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319.

Reddi, S. S., Rudin, S. F., and Keshavan, H. R. (1984). An optimal multiple threshold scheme for image segmentation. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-14(4):661–665.

Resnick, S. (1987). *Extreme Values, Regular Variation, and Point Processes*. Applied probability : a series of the applied probability trust. Springer-Verlag.

Ripley, B. D. (1996). *Belief Networks*, page 243–286. Cambridge University Press.

Rootzén, H. and Tajvidi, N. (2006). Multivariate generalized Pareto distributions. *Bernoulli*, 12(5):917–930.

Röttger, F., Engelke, S., and Zwiernik, P. (2023). Total positivity in multivariate extremes. *The Annals of Statistics*, 51(3):962–1004.

Roy, S., Atchadé, Y., and Michailidis, G. (2017). Change point estimation in high dimensional markov random-field models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1187–1206.

Santos, J. M. and Embrechts, M. (2009). On the use of the adjusted rand index as a metric for evaluating supervised classification. In Alippi, C., Polycarpou, M., Panayiotou, C., and Ellinas, G., editors, *Artificial Neural Networks – ICANN 2009*, pages 175–184, Berlin, Heidelberg. Springer Berlin Heidelberg.

Strokorb, K. (2020). Extremal independence old and new.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371.

Vignotto, E. and Engelke, S. (2020). Extreme value theory for anomaly detection – the gpd classifier. *Extremes*, 23:1–20.

Viktorov, I. and Abramov, A. (2020). The 2014–15 financial crisis in russia and the foundations of weak monetary power autonomy in the international political economy. *New Political Economy*, 25:487–510.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 71(3):671–683.

Wang, L., Ren, X., and Gu, Q. (2016). Precision matrix estimation in high dimensional gaussian graphical models with faster rates. In Gretton, A. and Robert, C. C., editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 177–185, Cadiz, Spain. PMLR.

A. Notation Schemes

Symbol	Explanation	Example
$[n]$	Denotes the set $\{1, \ldots, n\}$	
$[n_1 : n_2]$	Denotes the set $\{n_1, n_1 + 1, \ldots, n_2 - 1, n_2\}$	
$\mathcal{G}$	Denotes an (undirected) graph	$\mathcal{G} = (V, E)$
V	Denotes the set of nodes of $\mathcal{G}$	$V = [d]$
E	Denotes the set of (undirected) edges of $\mathcal{G}$	
i	Indicates temporal indexing of samples	$(\boldsymbol{y}_i)_{i \in [n]}$
$\cdot'$	Indicates a transformed temporal index	i'
$\mathcal{I}$	Denotes the non-transformed temporal index set	$\mathcal{I} = [n]$
m	Indicates spatial indexing across dimensions	$m \in [d]$
k, b	Indicates indexing across nodes of $\mathcal{G}$	$k, b \in V$
$\setminus k$	Denotes the set $[\lvert\mathcal{G}\rvert] \setminus \{k\}$	$b \in \setminus k$
$\lvert \cdot \rvert$	Denotes the cardinality of a set, or absolute value	$\lvert\mathcal{G}\rvert$ or $\lvert\beta_{b,i'}\rvert$
$\lVert \cdot \rVert_p$	Denotes the L_p-norm	$\lVert\boldsymbol{\beta}_{\cdot,i'}\rVert_2$
$\lfloor \cdot \rfloor$	Denotes the floor operator	$\lfloor \pi \rfloor = 3$
$\boldsymbol{\beta}_{\cdot,i'}$	Denotes the vector of regression coefficients for sample i'	
$\cdot^{(m)}$	Indicates association with the m-th transformation	$n^{(m)}$
$\cdot^{(k)}$	Indicates the k-th node acting as the response	$\hat{\boldsymbol{\beta}}^{(k,m)}$
N_C	Denotes the number of underlying change points	
$\mathcal{T}$	Denotes the set of underlying change points	
t_p	Denotes the index of the p-th change point	$\mathcal{T} = \{t_1, \ldots, t_{N_C}\}$
P_α	Denotes the α-th partition of the temporal index set	$P_1 = [t_1]$
$a_n = \mathcal{O}(b_n)$	$\lvert a_n \rvert$ asymptotically bounded above by b_n	
$a_n = \Theta(b_n)$	a_n asymptotically bounded above and below by b_n	
${}^rC_s$	Denotes the binomial coefficient of integers r and s	
$\times, \otimes$	Denotes the Cartesian product, and product σ-algebra	
$\xrightarrow{D}, \xrightarrow{\mathbb{P}}$	Denotes convergence in distribution, and probability	
$\mathcal{K}$	Denotes the kernel used for smoothing	
$\mathcal{K} * \boldsymbol{S}$	Denotes the convolution between $\mathcal{K}$ and vector $\boldsymbol{S}$	
$\tau, \tau_{\min}$	Denotes a threshold, and minimum threshold	
κ	Denotes the change sensitivity index	
$\Delta_{\min}$	Denotes the minimum number of edge changes	
δ_α	Denotes the number of edge changes from $\mathcal{G}_{\alpha-1}$ to $\mathcal{G}_\alpha$	
$dt_{\min}$	Denotes minimum separation between change points	
$\perp\!\!\!\perp$	Denotes regular (non-extremal) independence	
$\perp_e$	Denotes extremal independence	

Table 2.: Summary of notation used throughout the paper, similar to Kolar and Xing (2012).

B. Hüsler-Reiss Distributions and Gaussianity

In this section, we not only provide a Proof of Lemma 3.1, but also explicitly illustrate the transformation in Figure 5.

B.1. Proof of Lemma 3.1

Given that $\mathbf{Y} \sim \mathrm{HR}(\Gamma)$, where $\mathbf{Y} \in \mathbb{R}^d$, we aim to show that

$$\tilde{\mathbf{Y}}_{(m)} = \log\left(\mathbf{Y}^{(m)}_{\setminus m}\right) - \log\left(\mathbf{Y}^{(m)}_{m}\right) \sim \mathcal{N}\left(-\frac{\operatorname{diag}\left(\Sigma^{(m)}\right)}{2},\ \Sigma^{(m)}\right),$$

where recall $\mathbf{Y}^{(m)} = (\mathbf{Y} \mid Y_m > 1)$, with $m \in [d]$.

By Definition 3.7, we have that

$$f_{\mathbf{Y}}(\mathbf{y}) \propto y_m^{-2}\left(\prod_{i\neq m} y_i^{-1}\right)\varphi_{d-1}\left\{\log\left(\frac{\boldsymbol{y}_{\setminus m}}{y_m}\right) + \frac{\Gamma_{\setminus m,m}}{2},\ \Sigma^{(m)}\right\},$$

and so we have that

$$f_{\mathbf{Y}^{(m)}}(\boldsymbol{y}) = \frac{f_{\mathbf{Y}}(\boldsymbol{y})}{\mathbb{P}(Y_m > 1)} \propto y_m^{-2}\left(\prod_{i\neq m} y_i^{-1}\right)\varphi_{d-1}\left\{\log\left(\frac{\boldsymbol{y}_{\setminus m}}{y_m}\right) + \frac{\Gamma_{\setminus m,m}}{2},\ \Sigma^{(m)}\right\}.$$

In light of this, we consider the change of variables, $\boldsymbol{z} = \log(\boldsymbol{y})$, whose Jacobian determinant, and corresponding absolute value, is given by

$$\det J = \exp\left(\sum_{i=1}^{d} z_i\right) \quad \Rightarrow \quad |\det J| = \exp\left(\sum_{i=1}^{d} z_i\right).$$

Furthermore, we note that

$$y_m^{-2} = \exp\left(\log\left(y_m^{-2}\right)\right) = \exp\left(-2\log\left(y_m\right)\right) = \exp\left(-2z_m\right),$$

as well as

$$\left(\prod_{i\neq m} y_i^{-1}\right) = \exp\left(-\sum_{i\neq m} z_i\right),$$

and also that

$$\log\left(\frac{\boldsymbol{y}_{\setminus m}}{y_m}\right) = \log\left(\boldsymbol{y}_{\setminus m}\right) - \log\left(y_m\right)\cdot \mathbb{1} = \boldsymbol{z}_{\setminus m} - z_m \cdot \mathbb{1} = \tilde{\boldsymbol{z}}.$$

Thus, we have that

$$
\begin{aligned}
f_{\tilde{\mathbf{Z}},Z_m}(\tilde{\boldsymbol{z}}, z_m) &= f_{\mathbf{Y}^{(m)}}(\boldsymbol{y})|\det J| \\
&\propto \exp(-2z_m)\exp\left(-\sum_{i\neq m} z_i\right)\exp\left(\sum_i z_i\right)\varphi_{d-1}\left\{\tilde{\boldsymbol{z}} + \frac{\Gamma_{\setminus m,m}}{2},\ \Sigma^{(m)}\right\} \\
&= \exp(-z_m)\,\varphi_{d-1}\left\{\tilde{\boldsymbol{z}} + \frac{\Gamma_{\setminus m,m}}{2},\ \Sigma^{(m)}\right\}.
\end{aligned}
$$

In order to obtain the marginal density of $\tilde{\mathbf{Z}}$, we integrate over z_m,

$$
\begin{aligned}
f_{\tilde{\boldsymbol{z}}}(\tilde{\mathbf{Z}}) &= \int_0^\infty f_{\tilde{\mathbf{Z}},Z_m}(\tilde{\boldsymbol{z}}, z_m)dz_m \\
&\propto \left(\int_0^\infty \exp(-z_m)\,dz_m\right)\cdot\varphi_{d-1}\left\{\tilde{\boldsymbol{z}} + \frac{\Gamma_{\setminus m,m}}{2},\ \Sigma^{(m)}\right\} \\
&= \varphi_{d-1}\left\{\tilde{\boldsymbol{z}} + \frac{\Gamma_{\setminus m,m}}{2},\ \Sigma^{(m)}\right\},
\end{aligned}
$$

where we note that as we conditioned on $Y_m > 1$, we necessarily have that $z_m > 0$.

Finally noting that $\operatorname{diag}\left(\Sigma^{(m)}\right) = \Gamma_{\setminus m,m}$, we conclude that

$$
\tilde{\mathbf{Z}} \sim \mathcal{N}\left(-\frac{\operatorname{diag}\left(\Sigma^{(m)}\right)}{2},\ \Sigma^{(m)}\right),
$$

giving the desired result.

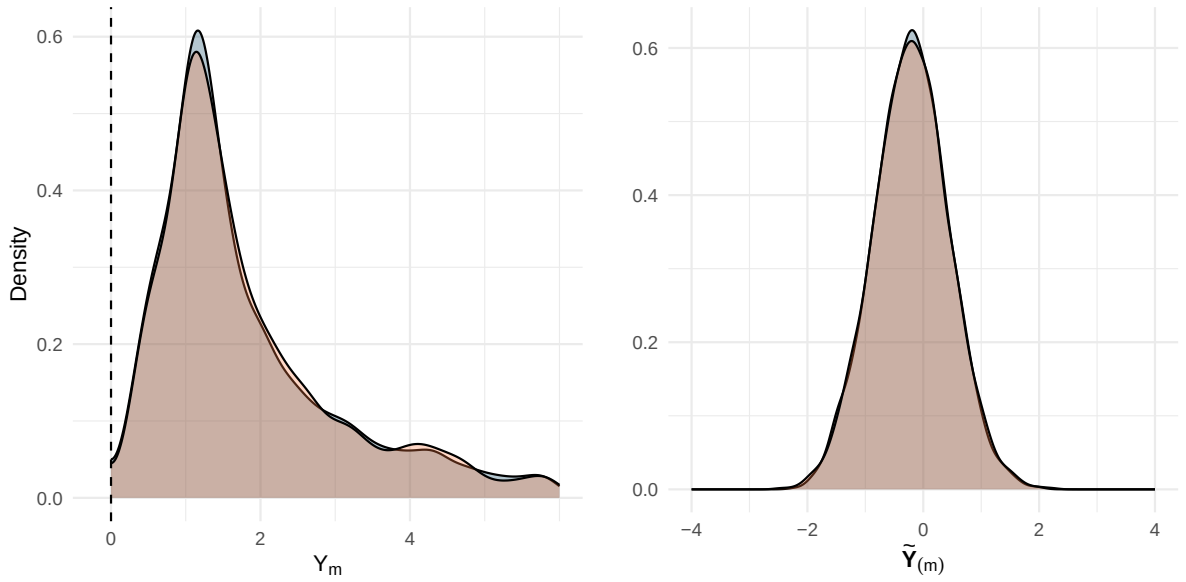

Figure 5.: Kernel density estimates of Y_m (left) and $\tilde{\mathbf{Y}}_{(m)}$ (right), for $m = 1$ (blue) and $m = 2$ (orange), where the original samples, $\mathbf{Y}$, correspond to those of Figure 2, which recall has support $\mathcal{L} = [0,\infty)^2 \setminus [0,1]^2$.

C. The Exponent Measure

In this section, we introduce the exponent measure, an example of a *radon* measure, that crucially enables us to outline how identical information pertaining to the tail of the random vector, $\boldsymbol{X}$, is recovered by the block maxima and POT approaches. In particular, we present the intuitive accounts outlined in Engelke and Hitz (2020) and Engelke and Ivanovs (2021), whilst refer the reader to Proposition 5.8 of Resnick (1987) for an alternative presentation.

A widely utilised assumption is that of *multivariate regular variation*, which specifically requires the existence of a Borel measure, Λ, defined on the space $\mathcal{E} = [0, \infty)^d \setminus \{0\}$, such that for all Λ-continuous Borel sets, $B \subset \mathcal{E}$, bounded away from the origin, this measure is finite, and the following limit exists,

$$\lim_{t \to \infty} t \cdot \mathbb{P}\left(\frac{\boldsymbol{X}}{t} \in B\right) = \Lambda(B). \tag{53}$$

If $\boldsymbol{X}$ is assumed to have standard Pareto margins, then from the block maxima perspective, Equation 13 simplifies to $\frac{\boldsymbol{M}_n}{n} \xrightarrow{D} \boldsymbol{Z}$, where $\boldsymbol{Z}$ is a max-stable random vector, and the marginals of $\boldsymbol{Z}$ are standard Fréchet (Engelke and Ivanovs, 2021), so that

$$\mathbb{P}(\boldsymbol{Z} \leq \boldsymbol{z}) = \exp\{-\Lambda_z\} = \exp\{-\Lambda(\mathcal{E} \setminus [0, \mathbf{z}])\}, \quad \mathbf{z} \in [0, \infty)^d, \tag{54}$$

and Λ is referred to as the exponent measure. Indeed, $\boldsymbol{X}$ is further said to be in the domain of attraction of the exponent measure, Λ (Engelke and Hitz, 2020).

On the other hand, from the perspective of the POT method, as outlined in Rootzén and Tajvidi (2006), as the threshold, $u \to \infty$, the corresponding exceedances converge in distribution to a MPD, $\boldsymbol{Y}$, such that we have

$$\mathbb{P}(\boldsymbol{Y} \leq \boldsymbol{z}) = \lim_{u \to \infty} \mathbb{P}\left(\frac{\boldsymbol{X}}{u} \leq \boldsymbol{z} \mid \|\boldsymbol{X}\|_\infty > u\right) = \frac{\Lambda_{\min(\boldsymbol{z}, \mathbf{1})} - \Lambda_{\boldsymbol{z}}}{\Lambda_{\mathbf{1}}}, \quad \boldsymbol{z} \in \mathcal{L}, \tag{55}$$

where specifically a componentwise minimum is taken in the exponent measure.

Crucially, when $\boldsymbol{X}$ is in the domain of attraction of the exponent measure, Λ, we equivalently have that $\boldsymbol{X}$ is in the domain of attraction of the max-stable distribution $\boldsymbol{Z}$, and the MPD, $\boldsymbol{Y}$.

D. Hill Plots

Analogous to the issue of block size selection in the block maxima method (Gumbel, 1958), the choice of threshold, u, above which observations are considered extreme, requires careful consideration in the POT approach. Indeed, a bias-variance trade-off corresponding to this choice is expected, as the asymptotic basis of the model is likely violated for too low a choice of u, whereas, for too high a choice of u, we expect high variability in model estimation due to an

insufficient number of exceedances (Coles, 2001). As discussed in detail within Coles (2001), this motivates the utilisation of stability analysis, of estimated parameters, to select the lowest value of u such that it is valid for corresponding exceedances to follow a GPD, which is notably preferred to *mean residual life plots* due to the common difficulty in interpreting the latter.

Specifically, the Hill estimator is a widely utilised estimate for the extreme value index, ξ^{-1}, of a heavy tailed distribution, such as the standard Pareto distribution, and is defined as

$$H_{k,n} = \frac{1}{k}\sum_{i=1}^{k} \log\left(\frac{X_{(i)}}{X_{(k+1)}}\right), \tag{56}$$

where $X_{(1)} \geq \cdots \geq X_{(n)}$ are the ordered observations, and $k+1$ is the total number of upper order statistics considered. As outlined in Drees et al. (2000), the above estimator is consistent for ξ^{-1}, that is $H_{k,n} \xrightarrow{\mathbb{P}} \xi^{-1}$, provided $\exists\,(k_n)_{n\in\mathbb{N}}$ such that $\frac{k_n}{n} \to 0$, as $k_n \to \infty$.

Practically, due to technical reasons discussed in the aforementioned paper, it is thus recommended to visualise the points, $\{(k, H_{k,n})\}_{k\in[n-1]}$, from which the value of ξ^{-1} is inferred from stability regions. Indeed, this plot is referred to as a *Hill plot*, and although can suffer from issues of interpretability when the underlying distribution deviates significantly from the Pareto distribution, this does provide a more robust approach than selecting an arbitrarily high quantile of the observed data.

E. Applications - Additional Details

E.1. Country Codes and $\hat{\mathcal{T}}$ for the Exchange Rate Dataset

Code	Foreign Exchange Rate (into GBP)	Code	Foreign Exchange Rate (into GBP)
AUS	Australian Dollar	POL	Polish Zloty
CHN	Chinese Yuan	RUS	Russian Ruble
DNK	Danish Krone	SWE	Swedish Krona
EUR	Euro	CHE	Swiss Franc
NOR	Norwegian Krone	USA	US Dollar

Table 3.: ISO 3166-1 alpha-3 country codes and the corresponding exchange rate (into GBP), for the subset of $d = 10$ currencies considered in Section 7.

Estimated Change Point	Corresponding Date	Nearby Major Currency Event
837	22 January 2009	End of 2007-2008 financial crisis
1128	17 March 2010	USD flash crash of 2010
1497	5 September 2011	SNB drops the EUR floor
2099	22 January 2014	Sanctions against Russia in 2014

Table 4.: Estimated change point indexes (in $\mathcal{I}$), corresponding dates, and related major currency events, as identified in Section 7.

E.2. Stability Plots of $\hat{\chi}_{\alpha\beta}$

In this section, we present a collection of stability plots for the estimated extremal correlation, $\hat{\chi}_{\alpha\beta}$, which were utilised in the application of the proposed procedure to a subset of foreign exchange rates in Section 7. Similarly to Hill plots, as described in Appendix D, examination of such plots amounts to achieving a reasonable balance between the bias and variance of the estimator, albeit in a visual manner. Notably, if one were to average multiple estimators together, the bias-variance trade-off may well be shifted due to the bias being affected differently to the variance in the corresponding aggregation. Therefore, individual consideration of each estimator is warranted, and note $k = \lfloor n^{0.75} \rfloor$ corresponds to stability regions in all cases.

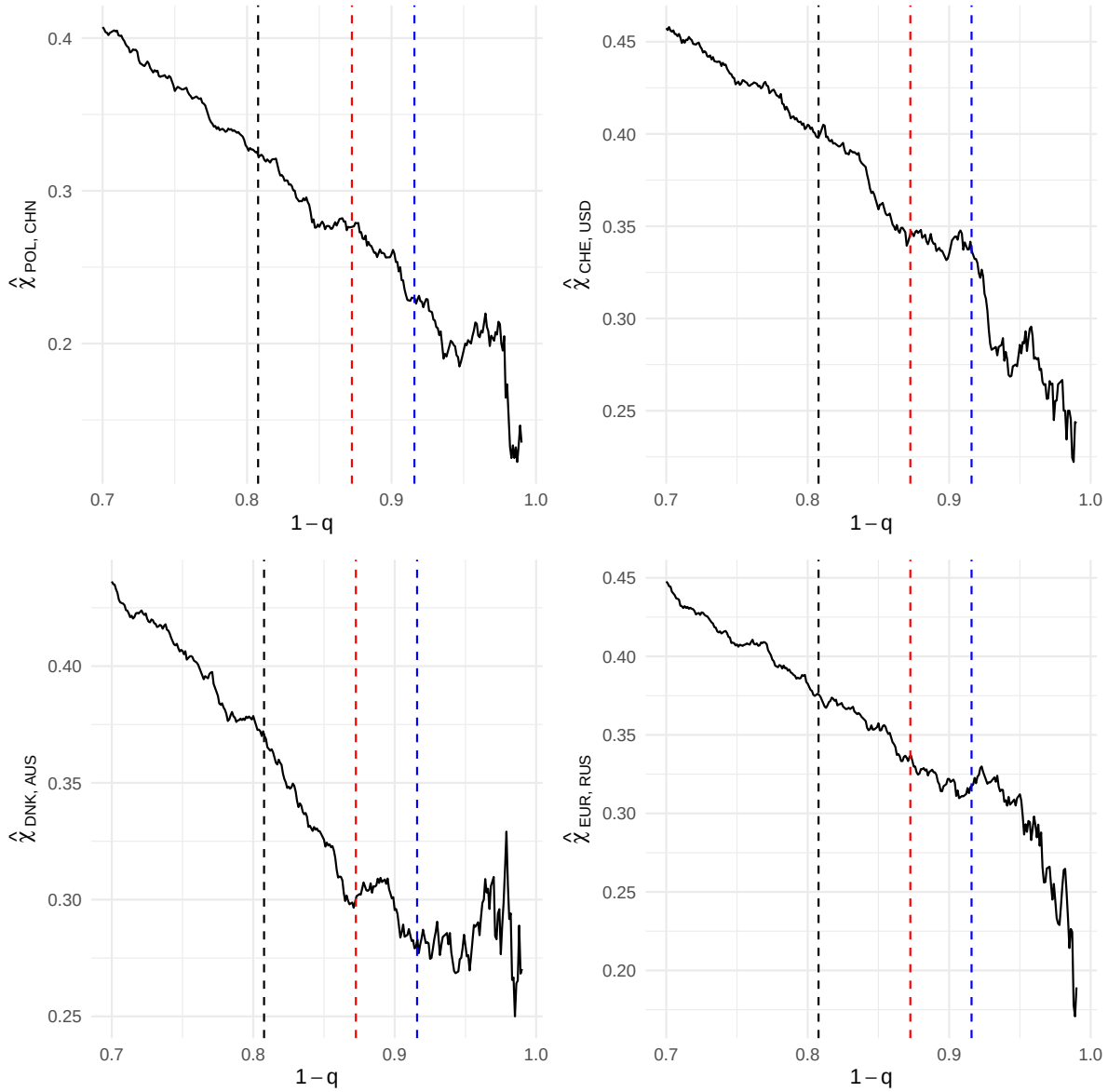

Figure 6.: Empirical estimates of the extremal correlation for a collection of exchange rates considered in Section 7. Values of $1 - q = 1 - \frac{k}{n}$, for $k = \lfloor n^{0.8} \rfloor$, $\lfloor n^{0.75} \rfloor$, and $\lfloor n^{0.7} \rfloor$, are indicated by black, red, and blue dashed lines respectively.

F. Simulation Studies - Results

d \ n	250	500	1000	2000
3	(0.00, 1.000, 1.000)	(0.00, 1.000, 1.000)	(0.00, 1.000, 1.000)	(0.00, 1.000, 1.000)
4	(0.00, 1.000, 1.000)	(0.00, 1.000, 1.000)	(0.00, 1.000, 1.000)	(0.00, 1.000, 1.000)
5	(0.00, 1.000, 1.000)	(0.00, 1.000, 1.000)	(0.00, 1.000, 1.000)	(0.00, 1.000, 1.000)
7	(0.00, 1.000, 1.000)	(0.00, 1.000, 1.000)	(0.00, 1.000, 1.000)	(0.00, 1.000, 1.000)
10	(0.02, 0.980, 1.000)	(0.08, 0.980, 1.000)	(0.00, 1.000, 1.000)	(0.00, 1.000, 1.000)
15	(0.20, 0.940, 1.000)	(0.14, 0.960, 1.000)	(0.00, 1.000, 1.000)	(0.00, 1.000, 1.000)
20	(0.20, 0.900, 1.000)	(0.24, 0.920, 1.000)	(0.00, 1.000, 1.000)	(0.00, 1.000, 1.000)

Table 5.: Average values of $(\Delta N_C, \text{ARI}, \text{ARI}^*)$, for $N_C = 0$ simulation study triplets, obtained over 50 simulations. Note that, in this case only, the ARI represents the percentage of simulations with no change point estimated, whereas $\text{ARI}^* = 1.0$ provided no change point was estimated in at least one simulation.

d \ n	250	500	1000	2000
3	(0.32, 0.694, 0.950)	(0.22, 0.801, 0.947)	(0.26, 0.729, 0.975)	(0.28, 0.715, 0.993)
4	(0.48, 0.772, 0.948)	(0.16, 0.908, 0.963)	(0.06, 0.955, 0.984)	(0.20, 0.861, 0.963)
5	(0.14, 0.923, 0.966)	(0.08, 0.939, 0.977)	(0.06, 0.953, 0.974)	(0.06, 0.961, 0.991)
7	(0.40, 0.909, 0.970)	(0.18, 0.920, 0.975)	(0.10, 0.940, 0.988)	(0.24, 0.934, 0.990)
10	(0.16, 0.946, 0.982)	(0.18, 0.917, 0.980)	(0.16, 0.950, 0.987)	(0.18, 0.929, 0.992)
15	(0.52, 0.874, 0.964)	(0.10, 0.951, 0.973)	(0.06, 0.941, 0.960)	(0.06, 0.963, 0.985)
20	(0.48, 0.851, 0.939)	(0.26, 0.914, 0.966)	(0.00, 0.986, 0.986)	(0.04, 0.959, 0.982)

Table 6.: Average values of $(\Delta N_C, \text{ARI}, \text{ARI}^*)$, for $N_C = 1$ simulation study triplets, obtained over 50 simulations.

d \ n	250	500	1000	2000
3	(0.94, 0.579, 0.711)	(0.90, 0.601, 0.827)	(1.00, 0.627, 0.868)	(0.72, 0.809, 0.984)
4	(0.64, 0.760, 0.833)	(0.62, 0.767, 0.901)	(0.60, 0.820, 0.958)	(0.50, 0.875, 0.993)
5	(0.90, 0.736, 0.941)	(0.62, 0.765, 0.899)	(0.56, 0.848, 0.966)	(0.66, 0.864, 0.991)
7	(1.00, 0.771, 0.947)	(0.54, 0.846, 0.959)	(0.54, 0.853, 0.978)	(0.44, 0.922, 0.992)
10	(1.00, 0.783, 0.878)	(0.52, 0.844, 0.979)	(0.44, 0.871, 0.990)	(0.36, 0.910, 0.993)
15	(1.04, 0.788, 0.915)	(0.56, 0.841, 0.946)	(0.52, 0.890, 0.987)	(0.30, 0.926, 0.992)
20	(1.22, 0.738, 0.736)	(0.72, 0.855, 0.980)	(0.34, 0.935, 0.987)	(0.18, 0.954, 0.995)

Table 7.: Average values of $(\Delta N_C, \text{ARI}, \text{ARI}^*)$, for $N_C = 2$ simulation study triplets, obtained over 50 simulations.

G. Proposed Procedure - Illustrated Example

In this section we illuminate the proposed procedure of Section 5.2, through explicit visualisation of each step of Algorithm 1. In particular, we consider the case of $N_C = 2$, $d = 6$, and $n = 500$, where the sequence of exact HR samples are generated by the procedure described in Section 6.

To begin with, we visualise, in Figure 7, the HRGMs associated with the piece-wise constant sequence of extremal variograms, $(\Gamma_i)_{i\in[500]}$, before and after the 2 change points, which in this case are explicitly given by $\mathcal{T} = \{158, 237\}$. To be clear, the proposed procedure has no knowledge of either $\mathcal{T}$, or N_C.

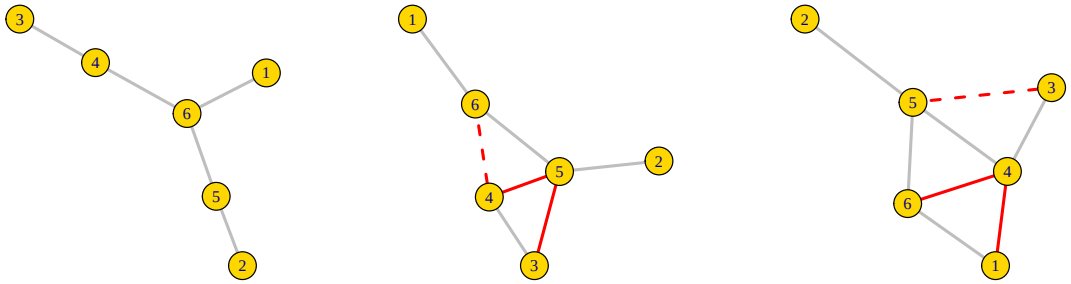

Figure 7.: The $d = 6$ HRGMs corresponding to the Γ_i matrices, before $t_1 = 158$, before $t_2 = 237$ and after $t_1 = 158$, and after $t_2 = 237$, from left to right. New extremal conditional dependence relationships are indicated by edges highlighted in red, whereas lost relationships are indicated through dashed red lines.

The sequence of exact HR samples, $(\boldsymbol{y}_i)_{i\in[500]} \in \mathbb{R}^6$, associated with $(\Gamma_i)_{i\in[500]}$, are visualised in Figure 8. By applying φ_m, 6 transformed sequences are obtained, namely $\left(\tilde{\boldsymbol{y}}_{i'}^{(m)}\right)_{i'\in[n^{(m)}]} \in \mathbb{R}^5$, for $m \in [6]$, and are visualised in Figure 9.

The regularisation parameter tuning procedure outlined in Section 5.2.3 is then performed, where we specifically tune each of $\lambda_1^{(m)}$ and $\lambda_2^{(m)}$ over 20 equally spaced values, between $[\frac{1}{2}\check{\lambda}_1^{(m)}, 2\check{\lambda}_1^{(m)}]$, and $[\frac{1}{2}\check{\lambda}_2^{(m)}, 2\check{\lambda}_2^{(m)}]$ respectively. The resulting estimates, obtained through utilising these selected regularisation parameters, are denoted by $\hat{\boldsymbol{\beta}}^{(k,m)}$, and for brevity, only $\hat{\boldsymbol{\beta}}^{(k,1)}$ values, for $k \in [5]$, are presented in Figure 10. We subsequently calculate the first differences, $\boldsymbol{\delta}^{(k,m)}$, and in the interest of continuity, present $\boldsymbol{\delta}^{(k,1)}$ in Figure 11. Next, we obtain the first difference sums, $\boldsymbol{S}^{(k,m)}$, presenting $\boldsymbol{S}^{(k,1)}$ in Figure 12, which note has been mapped from $\mathcal{I}'^{(m)}$ to $\mathcal{I}$. Then, we present $\tilde{\boldsymbol{S}}^{(m)}$, for $m \in [6]$, in Figure 13, which are further summed to obtain $\dot{\boldsymbol{S}}$, as presented in the upper panel of Figure 14. We finally calculate $\mathcal{K} * \dot{\boldsymbol{S}}$, where $\mathcal{K}$ is a 1D Gaussian distribution, with $\sigma = \frac{n}{100} = 5$, and τ is calculated using Otsu's method, as described in Section 5.2.2. In this case, we obtain $\hat{\mathcal{T}} = \{160, 237\}$, giving an ARI of 0.9922.

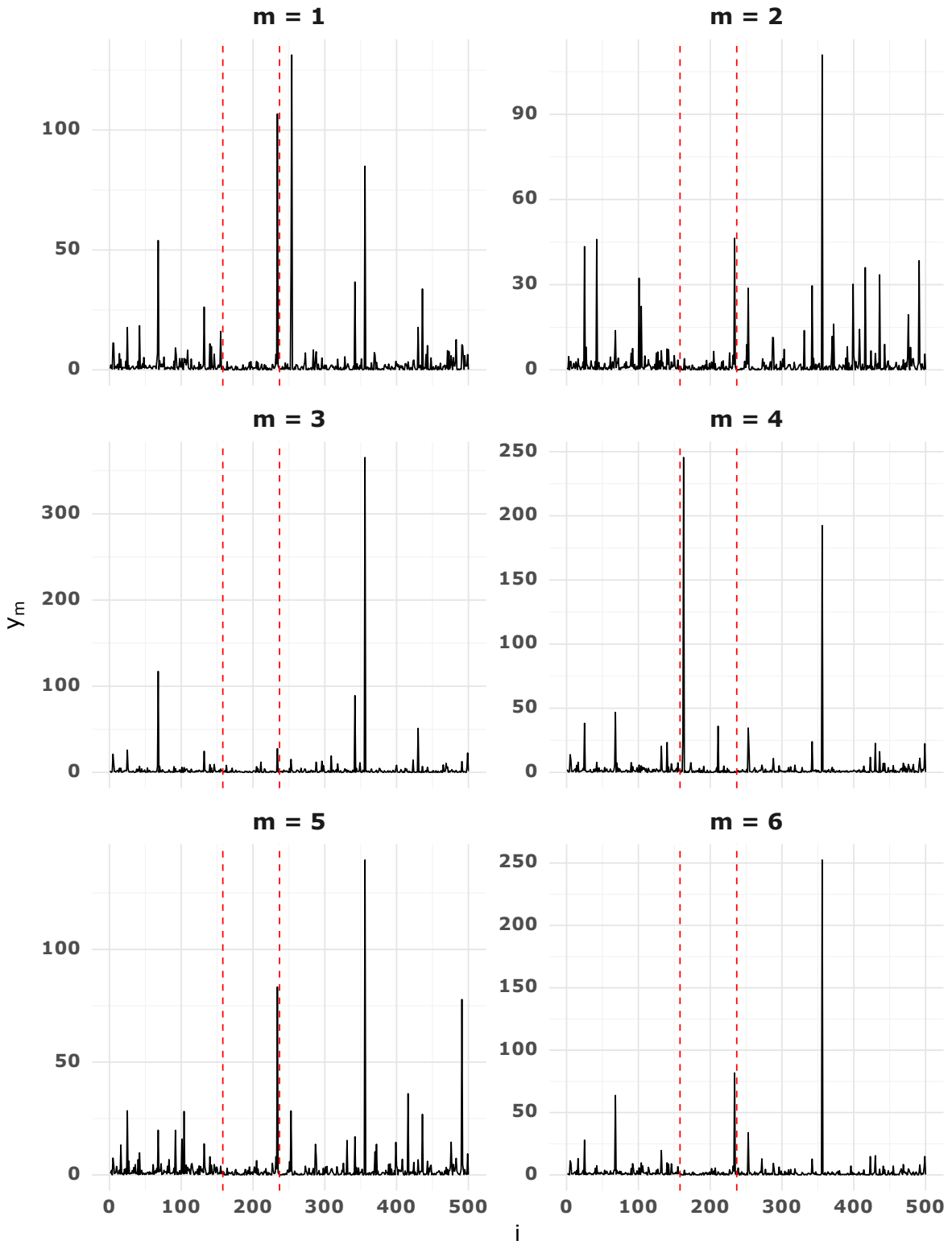

Figure 8.: Components of the exact HR samples generated for the illustrated example, using the Γ_i matrices associated with the HRGMs of Figure 7, where true change points are indicated by red dotted lines.

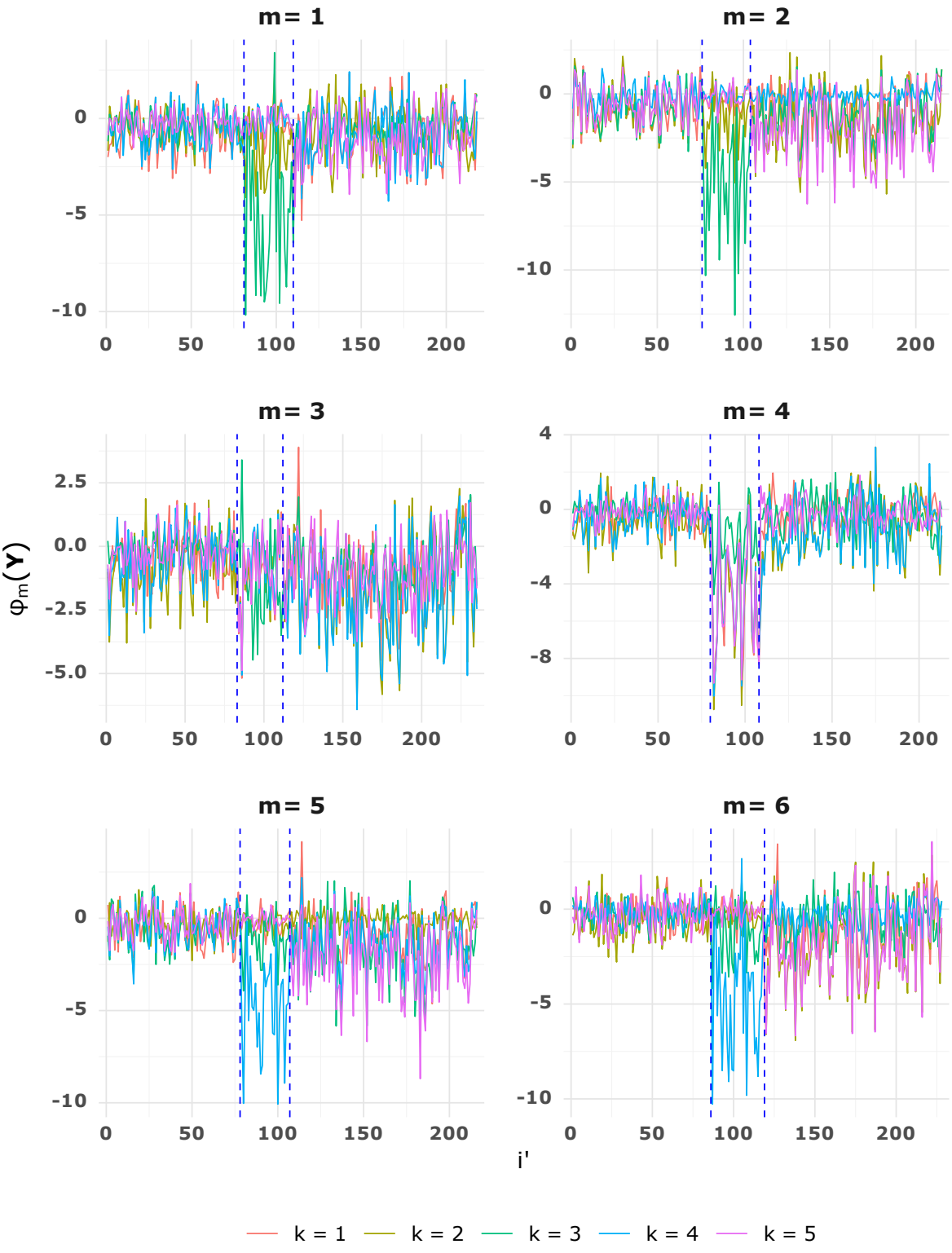

Figure 9.: The 6 sets of transformed HR samples, in the illustrated example, where φ_m is the transformation outlined in Lemma 3.1. Note, the blue dotted lines correspond to the optimal estimates of the true change points on $\mathcal{I}'^{(m)}$. Indeed, $n^{(m)} < n \; \forall m$, and $k \in [5]$.

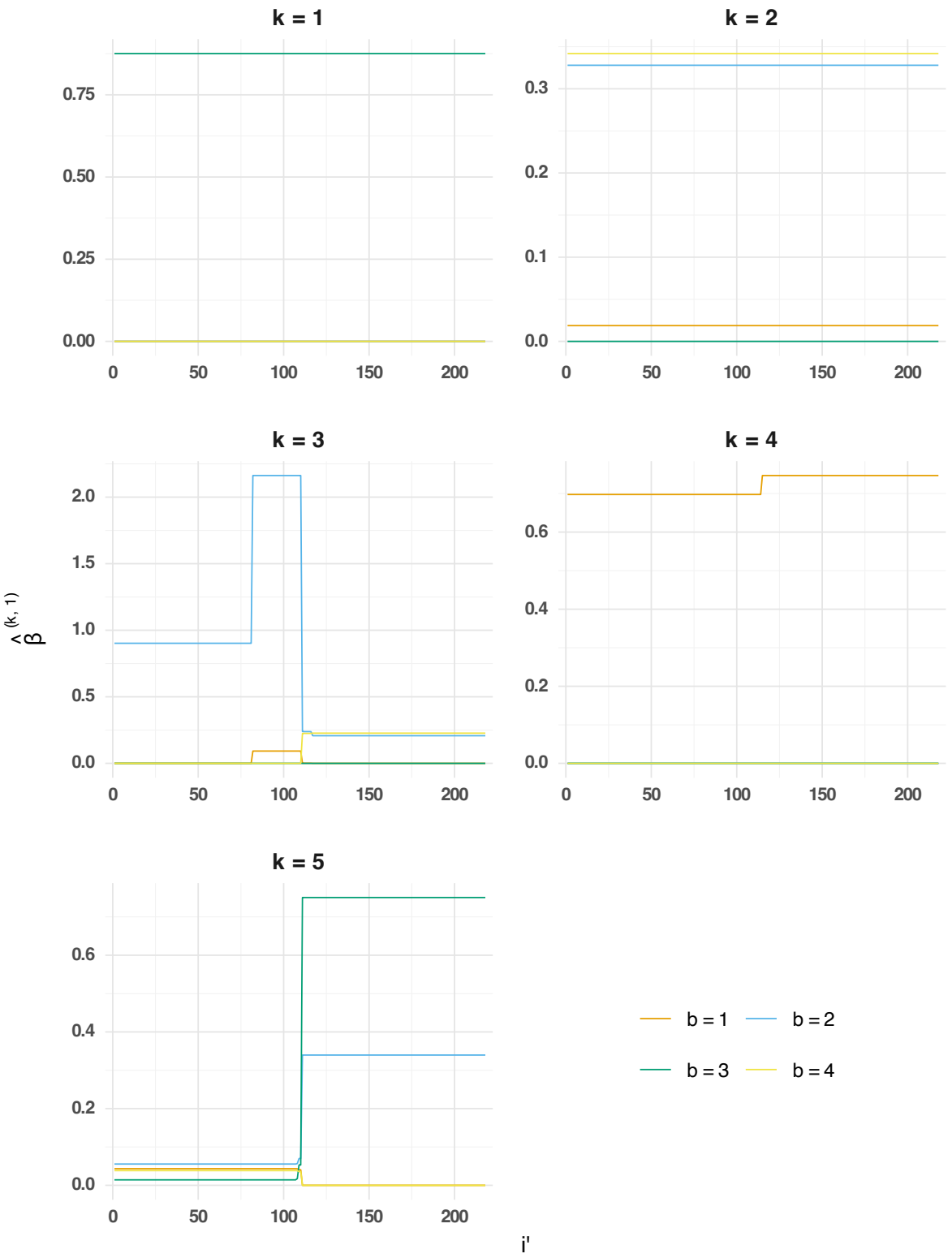

Figure 10.: The 5 sets of estimates, for the convex problem of Equation 30, using the $m = 1$ data from Figure 9, and the optimally tuned values of $\lambda_1^{(m)}$ and $\lambda_2^{(m)}$, as described in the illustrated example. Although omitted for brevity, the picture is similar for the other values of m.

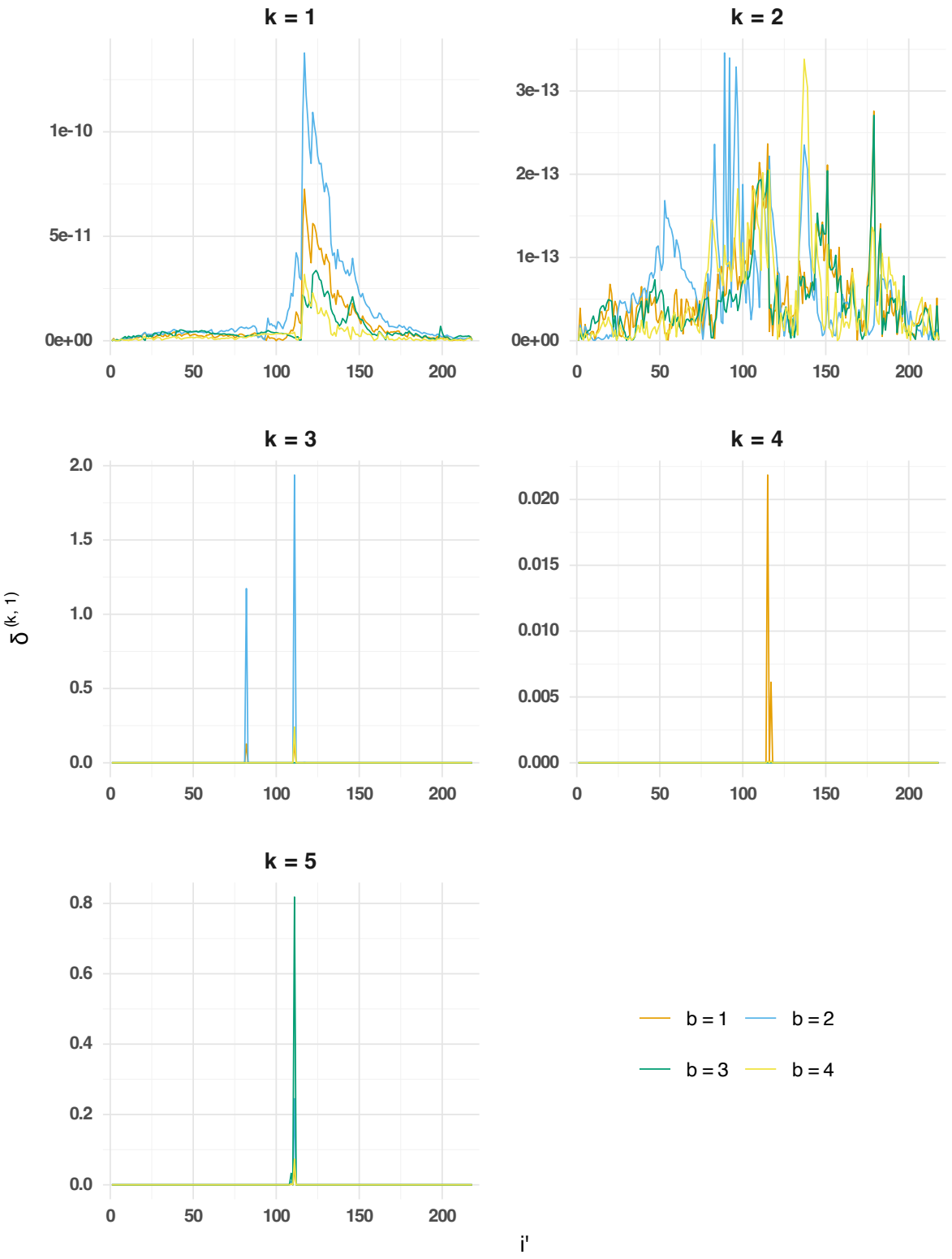

Figure 11.: The corresponding first differences for the estimates, $\beta^{(k,1)}$, of Figure 10. Notably, the $k = 1$ and $k = 2$ panels highlight the need for $\tau_{\min}$, as otherwise the observed noise would result in a significant number of false positives.

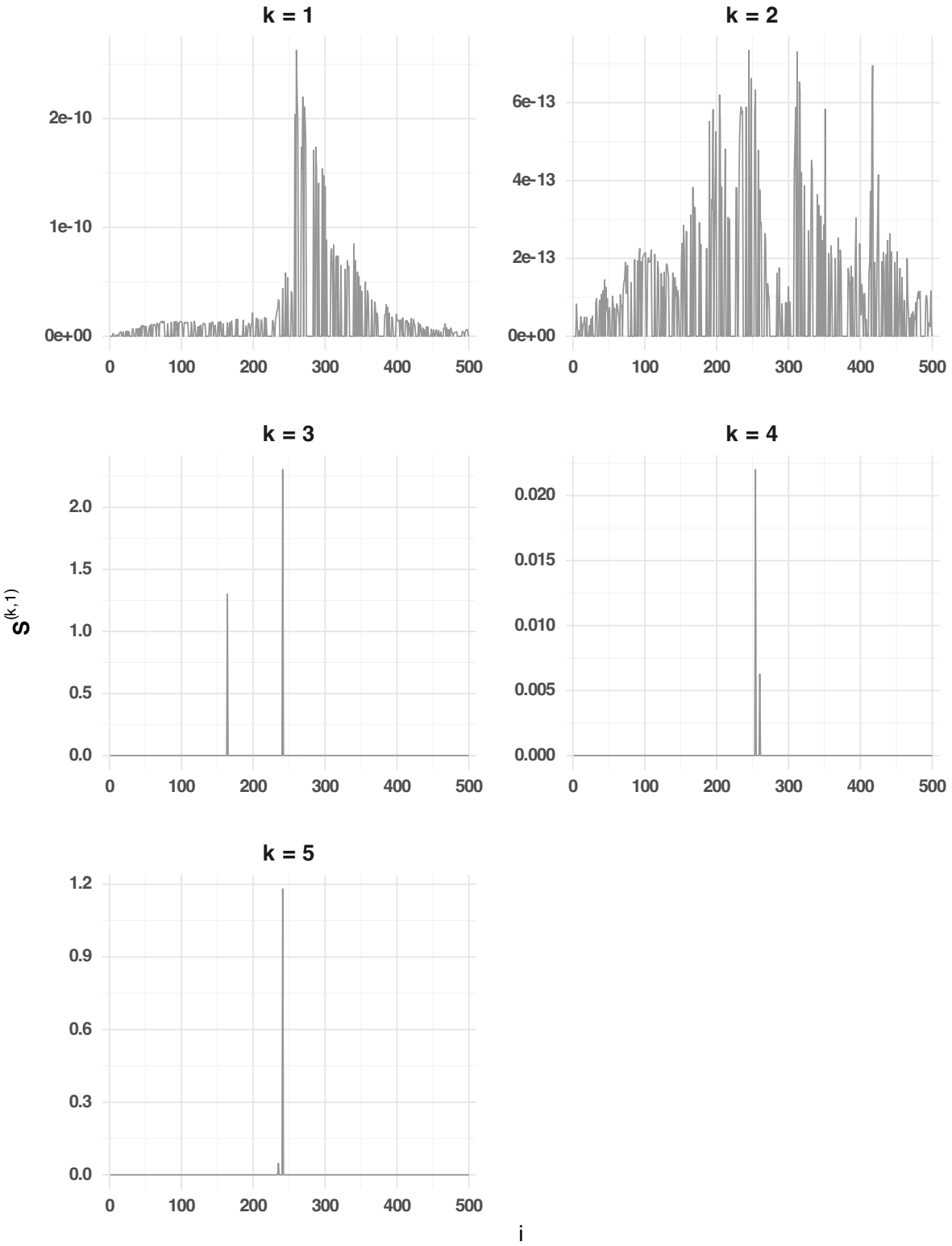

Figure 12.: The corresponding first difference sums, $\boldsymbol{S}^{(k,1)}$ of Figure 11, mapped from $\mathcal{I}'^{(1)}$ back to $\mathcal{I}$. The $k = 1$ and $k = 2$ panels are indicative of samples lost through conditioning utilised in φ_1, as they correspond to gaps in the curve for these cases.

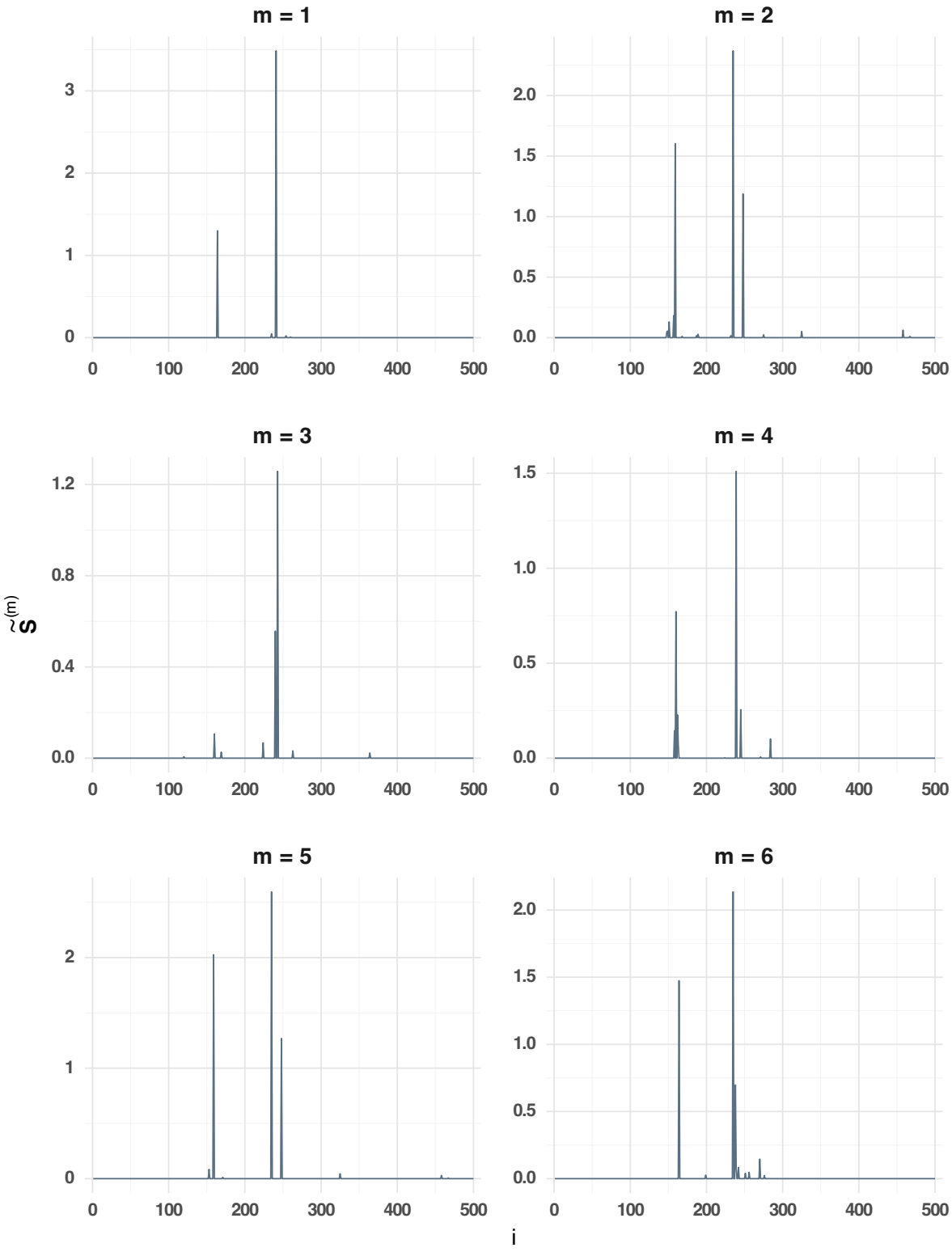

Figure 13.: The aggregated first difference sums, $\boldsymbol{S}^{(m)}$, for $m \in [6]$, of the illustrated example. Indeed, $\boldsymbol{S}^{(1)}$ is the result of the sum across the $\boldsymbol{S}^{(k,1)}$, for $k \in [5]$, as seen in Figure 12.

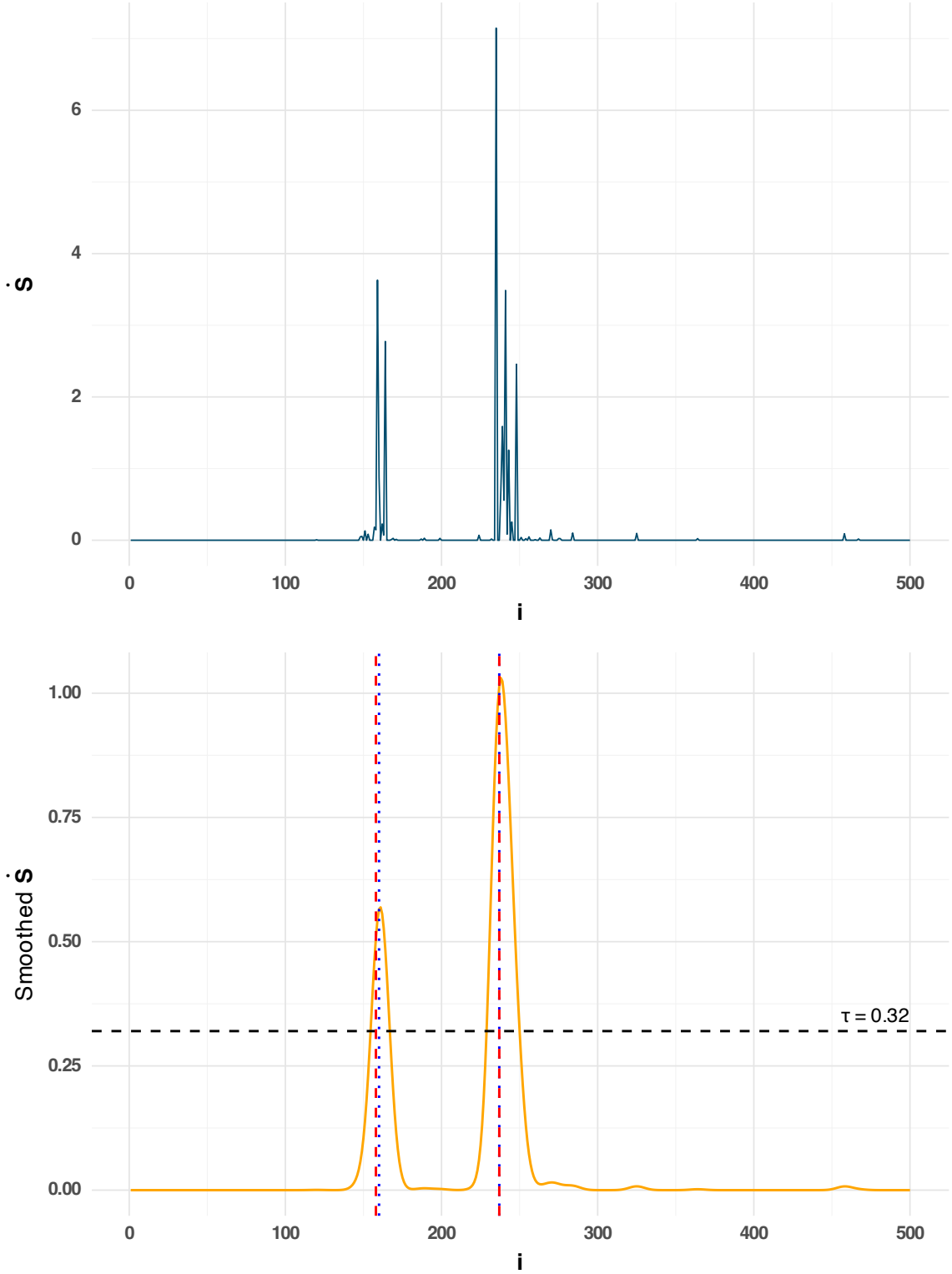

Figure 14.: The corresponding value of $\dot{\boldsymbol{S}}$ for the $\boldsymbol{S}^{(m)}$ of Figure 13 (top), and the smoothed version, $\mathcal{K} * \dot{\boldsymbol{S}}$ (bottom). The automatically selected threshold is indicated by the horizontal black dashed line, whereas the true and estimated change points are indicated by the vertical red dashed and blue dotted lines respectively.

H. Optimal Estimates - Technical Details

Recall, for each sequence of random vectors $(\tilde{\boldsymbol{y}}_{i'}^{(m)})_{i' \in [n^{(m)}]}$, we aim to solve the following estimation procedure for $k \in [d-1]$, and $\lambda_1, \lambda_2 \geq 0$:

$$\hat{\boldsymbol{\beta}}^{(k,m)} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{(d-2) \times n^{(m)}}}{\arg\min} \; \mathscr{L}(\boldsymbol{\beta}, k, m) + \mathscr{P}(\boldsymbol{\beta}, k, m, \lambda_1, \lambda_2), \tag{57}$$

where $\boldsymbol{\beta} = (\beta_{b,i'})$, with $b \in [d-2]$, $i' \in [n^{(m)}]$, and

$$\mathscr{L}(\boldsymbol{\beta}, k, m) = \sum_{i'=1}^{n^{(m)}} \left(\tilde{y}_{i',k}^{(m)} - \sum_{b \in \setminus k} \tilde{y}_{i',b}^{(m)} \beta_{b,i'} \right)^2, \tag{58}$$

$$\mathscr{P}(\boldsymbol{\beta}, k, m, \lambda_1, \lambda_2) = 2\lambda_1 \sum_{i'=2}^{n^{(m)}} \|\boldsymbol{\beta}_{\cdot,i'} - \boldsymbol{\beta}_{\cdot,i'-1}\|_2 + 2\lambda_2 \sum_{i'=1}^{n^{(m)}} \sum_{b \in \setminus k} |\beta_{b,i'}|. \tag{59}$$

H.1. Unrestricted Case - Proof of Lemma 5.1

The following is effectively the Proof of Lemma 1 in Kolar and Xing (2012) reformulated for the estimation procedure outlined above.

We first introduce a $(d-2) \times n^{(m)}$-dimensional matrix, $\boldsymbol{\gamma}^{(k,m)}$, defined as

$$\boldsymbol{\gamma}_{i'}^{(k,m)} = \begin{cases} \boldsymbol{\beta}_{\cdot,i'}^{(k,m)}, & \text{for } i' = 1, \\ \boldsymbol{\beta}_{\cdot,i'}^{(k,m)} - \boldsymbol{\beta}_{\cdot,i'-1}^{(k,m)}, & \text{otherwise,} \end{cases} \tag{60}$$

which notably means that the corresponding telescopic sum results in

$$\sum_{j' \leq i'} \gamma_{j',b}^{(k,m)} = (\beta_{b,i'}^{(k,m)} - \beta_{b,i'-1}^{(k,m)}) + (\beta_{b,i'-1}^{(k,m)} - \beta_{b,i'-2}^{(k,m)}) + \cdots + \beta_{b,1}^{(k,m)} = \beta_{b,i'}^{(k,m)}. \tag{61}$$

This crucially allows us to re-write the objective in (57) as

$$\begin{aligned} \hat{\boldsymbol{\gamma}}^{(k,m)} = \underset{\boldsymbol{\gamma} \in \mathbb{R}^{(d-2) \times n^{(m)}}}{\arg\min} & \sum_{i'=1}^{n^{(m)}} \left(\tilde{y}_{i',k}^{(m)} - \sum_{b \in \setminus k} \tilde{y}_{i',b}^{(m)} \sum_{j' \leq i'} \gamma_{j',b} \right)^2 \\ & + 2\lambda_1 \sum_{i'=2}^{n^{(m)}} \|\boldsymbol{\gamma}_{i'}\|_2 + 2\lambda_2 \sum_{i'=1}^{n^{(m)}} \sum_{b \in \setminus k} \left| \sum_{j' \leq i'} \gamma_{j',b} \right|, \end{aligned} \tag{62}$$

from which we remark that a necessary and sufficient condition for $\hat{\boldsymbol{\gamma}}^{(k,m)}$ to be a solution of (62), is that for each $l' \in [n^{(m)}]$, the $(d-2)$-dimensional zero vector, $\mathbf{0}$, belongs to the sub-differential of (62), with respect to $\boldsymbol{\gamma}_{l'}$, evaluated at $\hat{\boldsymbol{\gamma}}^{(k,m)}$. Indeed, this means that

$$\mathbf{0} = 2\sum_{i'=l'}^{n^{(m)}}(-\tilde{\boldsymbol{y}}^{(m)}_{i',\backslash k})\left(\tilde{y}^{(m)}_{i',k} - \sum_{b\in\backslash k}\tilde{y}^{(m)}_{i',b}\hat{\beta}^{(k,m)}_{b,i'}\right) + 2\lambda_1\hat{\boldsymbol{z}}_{l'} + 2\lambda_2\sum_{i'=l'}^{n^{(m)}}\hat{\boldsymbol{w}}_{i'}, \tag{63}$$

where $\hat{\boldsymbol{z}}_k \in \partial\|\cdot\|_2(\hat{\boldsymbol{\gamma}}_k)$, that is,

$$\hat{\boldsymbol{z}}_k = \begin{cases} \frac{\hat{\boldsymbol{\gamma}}_k}{\|\hat{\boldsymbol{\gamma}}_k\|_2}, & \text{if } \hat{\boldsymbol{\gamma}}_k \neq \mathbf{0}, \\ \in \mathcal{B}_2(0,1), & \text{otherwise}, \end{cases} \tag{64}$$

and $\boldsymbol{w}_i = \text{sign}(\sum_{j\leq i}\hat{\boldsymbol{\gamma}}_j)$ with $\text{sign}(\mathbf{0}) \in [-1,1]$. Lemma 5.1 then follows immediately.

H.2. Restricted Case - Proof of Lemma 5.2

Now, we first consider the restricted case, where for the sake of illustration, we suppose there is exactly one change point, occurring at $i' = t'_1 + 1$, such that $\left(\boldsymbol{\beta}^{(k,m)}_{i'}\right)_{i'\in[n^{(m)}]}$ is piece-wise constant about this change point. For clarity, we refer to the constant vector of $\left(\boldsymbol{\beta}^{(k,m)}_{i'}\right)_{i'\in[t'_1]}$ by $\boldsymbol{\theta}^{(0)}$, and the constant vector after the change point as $\boldsymbol{\theta}^{(1)}$. Furthermore, note that t'_1 is, if it exists, the index within $\mathcal{I}'^{(m)}$ such that $\phi_m(t'_1+1) = t_1$, and otherwise is index before the optimal estimate of t_1, in the sense described within Figure 3.

Note that now, we have

$$\boldsymbol{\gamma}^{(k,m)}_{i'} = \begin{cases} \boldsymbol{\theta}^{(0)}, & \text{for } i' = 1, \\ \mathbf{0}, & \text{for } 2 \leq i' \leq t'_1, \\ \boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}, & \text{for } i' = t'_1 + 1, \\ \mathbf{0}, & \text{for } t'_1 + 2 \leq i' \leq n^{(m)}, \end{cases} \tag{65}$$

Thus, we have that (57) can be re-written as

$$\begin{aligned} \hat{\boldsymbol{\gamma}}^{(k,m)} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^{(d-2)\times n^{(m)}}}{\arg\min} & \sum_{i'=1}^{t'_1}\left(\tilde{y}^{(m)}_{i',k} - \sum_{b\in\backslash k}\tilde{y}^{(m)}_{i',b}\gamma^{(k,m)}_{1,b}\right)^2 \\ & + \sum_{i=t'_1+1}^{n^{(m)}}\left(\tilde{y}^{(m)}_{i',k} - \sum_{b\in\backslash k}\tilde{y}^{(m)}_{i',b}(\gamma^{(k,m)}_{1,b} + \gamma^{(k,m)}_{t'_1+1,b})\right)^2 \\ & + 2\lambda_1\|\boldsymbol{\gamma}^{(k,m)}_{t'_1+1}\|_2 \\ & + 2\lambda_2\left(t'_1\sum_{b\in\backslash k}|\gamma^{(k,m)}_{1,b}| + (n^{(m)} - t'_1)\sum_{b\in\backslash k}|\gamma^{(k,m)}_{1,b} + \gamma^{(k,m)}_{t'_1+1,b}|\right), \end{aligned} \tag{66}$$

and notably sub-gradients are always equal to the zero vector in cases that $l' \neq 1, t'_1 + 1$.

Taking sub-gradients with respect to $\boldsymbol{\gamma}_1^{(k,m)}$, and equating to the zero vector results in

$$\begin{aligned}
\mathbf{0} = & 2\sum_{i'=1}^{t'_1}(-\tilde{\boldsymbol{y}}_{i',\backslash k}^{(m)})\left(\tilde{y}_{i',k}^{(m)} - \sum_{b\in\backslash k}\tilde{y}_{i',b}^{(m)}\hat{\theta}_b^{(0)}\right) \\
& + 2\sum_{i'=t'_1+1}^{n^{(m)}}(-\tilde{\boldsymbol{y}}_{i',\backslash k}^{(m)})\left(\tilde{y}_{i',k}^{(m)} - \sum_{b\in\backslash k}\tilde{y}_{i',b}^{(m)}\hat{\theta}_b^{(1)}\right) \\
& + 2\lambda_2 \cdot t'_1 \cdot \text{sign}(\hat{\boldsymbol{\gamma}}_1^{(k,m)}) \\
& + 2\lambda_2 \cdot (n^{(m)} - t'_1) \cdot \text{sign}(\hat{\boldsymbol{\gamma}}_1^{(k,m)} + \hat{\boldsymbol{\gamma}}_{t'_1+1}^{(k,m)}),
\end{aligned} \tag{67}$$

whereas with respect to $\boldsymbol{\gamma}_{t'_1+1}^{(k,m)}$, and equating to the zero vector, we obtain

$$\begin{aligned}
\mathbf{0} = & 2\sum_{i'=t'_1+1}^{n^{(m)}}(-\tilde{\boldsymbol{y}}_{i',\backslash k}^{(m)})\left(\tilde{y}_{i',k}^{(m)} - \sum_{b\in\backslash k}\tilde{y}_{i',b}^{(m)}\hat{\theta}_b^{(1)}\right) \\
& + 2\lambda_1 \frac{\hat{\boldsymbol{\gamma}}_{t'_1+1}^{(k,m)}}{\|\hat{\boldsymbol{\gamma}}_{t'_1+1}^{(k,m)}\|_2} \\
& + 2\lambda_2 (n^{(m)} - t'_1)\text{sign}(\hat{\boldsymbol{\gamma}}_1^{(k,m)} + \hat{\boldsymbol{\gamma}}_{t'_1+1}^{(k,m)}),
\end{aligned} \tag{68}$$

where we note that $\hat{\boldsymbol{\gamma}}_{t'_1+1} \neq \mathbf{0}$.

Notably, in this case, only $N_C + 1 = 2$ equations must be satisfied, whereas in the unrestricted case we have $n^{(m)}$ equations. Furthermore, note that $\hat{\theta}_b^{(0)} = \hat{\gamma}_{1,b}^{(k,m)}$, whilst $\hat{\theta}_b^{(1)} = \hat{\gamma}_{1,b}^{(k,m)} + \hat{\gamma}_{t'_1,b}^{(k,m)}$, and so the optimality equations can be written solely in terms of $\hat{\boldsymbol{\gamma}}^{(k,m)}$.

In the more general case, of N_C known change points at locations, $\mathcal{T} = \{t_1, \ldots, t_{N_C}\}$, and $\left(\boldsymbol{\beta}_{i'}^{(k,m)}\right)_{i'\in[n^{(m)}]}$ is piece-wise constant about the corresponding change points on the transformed temporal index set, $\mathcal{I}'^{(m)}$, we follow the same logic to obtain $N_C + 1$ optimality equations, where the optimality equation corresponding to the p-th change point is characterised by

$$\begin{aligned}
\mathbf{0} = & 2\sum_{i'=t'_p+1}^{t'_{p+1}}(-\tilde{\boldsymbol{y}}_{i',\backslash k}^{(m)})\left(\tilde{y}_{i',k}^{(m)} - \sum_{b\in\backslash k}\tilde{y}_{i',b}^{(m)}\hat{\theta}_b^{(p)}\right) \\
& + 2\sum_{i'=t'_{p+1}+1}^{t'_{p+2}}(-\tilde{\boldsymbol{y}}_{i',\backslash k}^{(m)})\left(\tilde{y}_{i',k}^{(m)} - \sum_{b\in\backslash k}\tilde{y}_{i',b}^{(m)}\hat{\theta}_b^{(p+1)}\right) \\
& + 2\lambda_1 \hat{\mathbf{z}}_{t'_p+1} \\
& + 2\lambda_2 \sum_{q=p}^{N_C}\hat{\mathbf{w}}_q,
\end{aligned} \tag{69}$$

where $t'_0 = 0$, $t'_{N_C+1} = n^{(m)}$, $\hat{\boldsymbol{\theta}}^{(N_C+1)} = \mathbf{0}$, and to be clear, t'_0 is considered the 0-th change point no matter the value of N_C, and forms the additional optimality equation. Somewhat similarly to the unrestricted case, we have that

$$\hat{\mathbf{z}}_{t'_p+1} = \begin{cases} \frac{\hat{\boldsymbol{\gamma}}_{t'_p+1}}{\|\hat{\boldsymbol{\gamma}}_{t'_p+1}\|_2}, & \text{if } \hat{\boldsymbol{\gamma}}_{t'_p+1} \neq \mathbf{0} \text{ and } p \neq 0, \\ \in \mathcal{B}_2(0,1), & \text{otherwise,} \end{cases} \tag{70}$$

and $\mathbf{w}_q = \text{sign}(\sum_{j \leq q} \hat{\boldsymbol{\gamma}}_j)$ with $\text{sign}(\mathbf{0}) \in [-1, 1]$. Lemma 5.2 then follows immediately. Indeed, in the case of $N_C = 1$, note that subbing in $p = 0$ into (69) recovers (67), whereas subbing in $p = 1$ into (69) recovers (68).